2010

Identification and Characterization of APOBEC1 mRNA Editing Targets: A Transcriptomics Approach

Brad Randall Rosenberg

IDENTIFICATION AND CHARACTERIZATION OF APOBEC1 MRNA EDITING TARGETS: A TRANSCRIPTOMICS APPROACH

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Brad Randall Rosenberg

June 2010

IDENTIFICATION AND CHARACTERIZATION OF APOBEC1 MRNA EDITING TARGETS: A TRANSCRIPTOMICS APPROACH

Brad Randall Rosenberg, Ph.D.
The Rockefeller University 2010

RNA editing is generally defined as the alteration of an RNA sequence from that encoded by the genome through nucleotide insertion, deletion or modification. The Apolipoprotein B mRNA Editing Catalytic polypeptide 1 (APOBEC1) cytidine deaminase is an mRNA editing enzyme that modifies a specific cytidine in the apolipoprotein B (apoB) transcripts of small intestine enterocytes. APOBEC1-mediated cytidine to uridine editing generates an in-frame stop codon and results in translation of a truncated apoB isoform with distinct functions in lipid transport. Other physiological mRNA targets of APOBEC1 editing have remained largely unknown.

This thesis presents the development of an RNA-Seq method for the identification of mRNA editing events on a transcriptome-wide scale and its application to the discovery of APOBEC1 editing targets in small intestinal enterocytes. The technique utilizes ultra-high throughput sequencing and subsequent bioinformatic analysis to compare wild-type and congenic *apobec1*$^{-/-}$ transcriptomes for specific single nucleotide variants indicative of editing.

Following technical validation, the screening approach was used to identify more than 30 previously undescribed APOBEC1 editing sites in enterocyte mRNA. All of the newly identified sites are located in AU-rich segments of transcript 3′ untranslated regions (3′ UTRs). Furthermore, these sites share several characteristic sequence features, including flanking nucleotide preferences and a downstream (3′) APOBEC1 mooring motif. Sequence pattern

recognition analysis based on these features successfully predicted additional APOBEC1 targets.

The studies detailed here demonstrate the feasibility and utility of a novel transcriptomics approach to RNA editing studies. The corresponding results indicate that APOBEC1 site-specifically edits many mRNA transcripts other than apoB in small intestinal enterocytes, suggesting additional roles for APOBEC1 beyond its function in apolipoprotein regulation.

ACKNOWLEDGMENTS

I am extraordinarily grateful for all of the instruction, assistance and support that I have received throughout my graduate work. First and foremost, I would like to thank my primary advisor, Nina Papavasiliou, for her limitless dedication to this project and to my scientific education as a whole. Nina has been a leader, a teacher, a guide, a colleague, and a friend. I cannot overstate my gratitude for her constant support.

Throughout my time in graduate school, I have been particularly fortunate to have not one, but two outstanding advisors. As such, I would like to acknowledge Charles Rice, who has been an exceptional mentor. No matter where my scientific pursuits have led me, Charlie has provided continuous guidance and encouragement.

I would also like to thank the members of my Faculty Advisory Committee, Christian Münz, Paul Bieniasz, and Carl Nathan, for their helpful advice and thoughtful suggestions. Despite different institutions and even different continents, their commitment has remained unwavering. In addition, I would like to express my gratitude to Juan Alfonzo from Ohio State University for serving as the external examiner at my thesis defense.

This work would not have been possible without Scott Dewell, who provided bioinformatics training, invaluable assistance, and lots and lots of short read sequences. I would also like to thank Michael Mwangi for his phylogenetics analyses, biostatistics contributions and helpful discussions.

It has been a privilege working with the current and former members of the Laboratory of Lymphocyte Biology, including Eva Besmer, Eleonora Market, Irena Pastar, Grace Teng, Tatyana Leonova, Peter "The Beast" Alff, Catharine Boothroyd, Jan Davidson-Moncada, Galadriel Hovel-Miner, Rebecca Delker, Ina Ly, Claire "Chams" Hamilton, Marianne Labriola, Messrs. Paul Hakimpour, Alexandros Strikoudis, and Eric "Fritz" Fritz. In particular, I would like to acknowledge Claire, who has been an instrumental part of this project and a pleasure to work with. I would also like to thank the members of the Laboratory of Molecular Virology for sharing their expertise and always making me feel welcome.

I would like to thank Thomas Tuschl, Volker Hovestadt and Aleks Mihailovic for microRNA analysis. I am also grateful to Robert Darnell for his guidance and support, and to the members of his laboratory for their constant willingness to help. In particular, I would like to acknowledge Donny Licatalosi, Jennifer Darnell, Nathalie Blachere, Randy Longman, Joseph Luna, Sarah van Driesche and Emily Conn.

My time in the lab was greatly enriched by the company and friendship of numerous colleagues. I would like to thank Robert Anthony, Kate Jeffrey, Vanessa Bryant, Reuben Richards, Patrick Smith, Jose Pagan, Rene Ott and Falk Nimmerjahn for providing scientific assistance and making every day an interesting and enjoyable one.

I would also like to acknowledge Matthew Albert for providing guidance and many unique scientific opportunities. Matthew is a valued mentor and role model, and I'm very grateful for his support and friendship.

Both The Rockefeller University Dean's Office and the Tri-Institutional MD-PhD Program provided tireless and unfailing assistance throughout this process. A special acknowledgment is due to Olaf Andersen for his guidance and his dedication to the training of physician scientists.

I owe a tremendous debt of gratitude to all of my friends and loved ones who have supported me through the ups and downs that come with scientific research. In particular, I would like to express my sincerest love and appreciation to my "core four": Katrina McGinty, Robert McGinty, Joanna Spencer and Ernesto Gonzalez. No one could ask for better friends. I would also like to acknowledge Genia Livshits, who has provided incredible personal and scientific support.

Most importantly, I would like to thank my immediate and extended family. I consider myself extraordinarily fortunate for all that they have provided throughout my life. Finally, a titanic acknowledgment goes to my parents, who have supplied boundless support, encouragement and love.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 1: INTRODUCTION

An organism's genome contains the complete set of information required for its development, function, survival and reproduction. This exceedingly complex biological compilation is encoded entirely by the four nucleotides of DNA: adenine, cytosine, guanine and thymine. Changes in nucleotide sequence can directly impact the information content encoded therein, with corresponding functional consequences. As the properties of DNA, RNA and proteins are determined by their sequences, molecular biology can be thought of as an information science. All life is dependent on the maintenance, interpretation and replication of vast amounts information represented by linear biopolymers composed of simple components.

The human genome contains approximately 25,000 genes (Consortium, 2004). This is only about twice that of the fruit fly (Adams et al., 2000; Misra et al., 2002) and four times that of the baker's yeast (Goffeau et al., 1996; Mewes et al., 1997). Considering the complexity of the human organism, this number seems surprisingly small. As fixed units, it is unlikely that 25,000 genes could allow for sufficient informational content to represent the complexity of mammalian life. Therefore, numerous mechanisms exist to impart additional diversity and complexity to the genome, transcriptome and proteome. Programmed changes in genomic DNA are rare, likely due to its heritability and the potentially deleterious effects of off-target errors. However, immune cells rely on somatic recombination and mutation to diversify their antigen receptors. In contrast, proteins are subject to extensive post-translational modifications that can have long- and/or short-term impact on their functionality. Changes at the

RNA level, however, represent perhaps the most significant source of complexity in the content and regulation of the expressed proteome.

As transitory reproductions of information permanently "hard coded" in the genome, RNA transcripts can undergo many types of diversity-promoting alterations without compromising the genetic integrity of the cell. Alternative mRNA splicing generates a dramatic increase in transcriptome complexity. Most mammalian genes contain multiple exons and more than 75% are estimated to have at least 2 alternatively spliced transcript isoforms (Kampa et al., 2004). Through the combinatorial generation of different mature transcript forms, alternative splicing increases the number of unique "functional genes" (i.e. distinct proteins) by a considerable factor. Additionally, other post-transcriptional processing events can alter start codon usage, affecting the length or reading frame of translated proteins. Differential poly-adenylation patterns can influence transcript stability and regulation.

Though all of these processes alter the content of RNA transcripts, they do so by including or excluding sequence blocks encoded by genomic DNA without modifying the individual nucleotides. Additional diversity is imparted by RNA editing processes, in which the primary sequence of an RNA transcript is modified by chemical changes to its nucleotide content. As such, RNA editing is a flexible mechanism for introducing very precise alterations to the transcriptome at single nucleotide resolution. Such seemingly minor changes to nucleotide sequence can have functional consequences ranging from the subtle to the dramatic in a variety of biological contexts.

1.1. RNA Editing – An Overview

RNA editing is generally defined as the co- or post-transcriptional alteration of an RNA sequence from that encoded by the genome through nucleotide insertion, deletion or modification (Wedekind et al., 2003). Though not all editing mechanisms are evolutionarily related, RNA editing has been observed in all domains of life: Archaea, Bacteria and Eukarya. The substrates and functional outcomes of editing are extremely varied; editing can impact RNA structure, base pairing, and RNA-protein interactions. In mRNA, editing can create, delete or alter codons, introduce novel splice sites, and modulate transcript stability and regulation (Bass, 2002).

1.1.1. Uridine Insertion/Deletion Editing

The two primary mechanisms of RNA editing are uridine insertion/deletion editing and base modification editing. Uridine insertion/deletion has been observed in the mitochondrial gene transcripts of the kinetoplastids, including *Trypanosoma brucei* and *Leishmania tarantolae*. Mitochondrial gene expression by these flagellated protozoa requires a series of complex RNA processing events, which for many genes involves the insertion and/or deletion of uridine residues within the transcripts' primary sequence. Initial examinations of different kinetoplastid mitochondrial gene sequences revealed aberrant cryptogenes: sequences with similarity to essential mitochondrial genes but distorted with frameshift mutations (Benne et al., 1986), missing start codons and/or tracts of absent coding sequence (Simpson et al., 1987). Following the recognition that *T. brucei* cytochrome oxidase subunit II (*coxII*) mRNA contains additional uridines that effectively "repair" a genome-

encoded frameshift (Benne et al., 1986), RNA editing was recognized as an important mechanism for functional gene expression.

Uridine insertion/deletion editing in kinetoplastids is a complex process mediated by numerous protein and RNA components. The information for editing at particular sites is specified by guide RNAs (Blum et al., 1990; Sturm and Simpson, 1990b), most of which are encoded by kinetoplastid minicircle DNA (Sturm and Simpson, 1990a). Based on sequence complementarity in their 5′ anchor region, guide RNAs form heterohybrids with target mRNAs and provide a template for the insertion (Blum et al., 1990; Sturm and Simpson, 1990b) or deletion (Seiwert and Stuart, 1994) of uridines by Watson-Crick and G:U base-pairing rules. Nucleotide insertion or deletion is performed by different multimeric protein complexes (Schnaufer et al., 2003), which contain RNA ligase, U-specific exonuclease, site-specific endonuclease, uridylyl transferase and numerous RNA binding proteins (Aphasizhev et al., 2003; Panigrahi et al., 2003a; Panigrahi et al., 2003b; Simpson et al., 2004; Stuart et al., 2004). Editing has been observed in more than half of kinetoplastid mitochondrial mRNAs (Feagin et al., 1988; Shaw et al., 1989; Shaw et al., 1988) and is essential for kinetoplastid survival (Schnaufer et al., 2001).

Uridine insertion/deletion RNA editing is unique to kinetoplastids. Given the complexity and considerable metabolic cost of this process, it remains unclear as to why this editing mechanism was selected for and maintained in these organisms. One theory speculates that editing evolved to repair mutations accumulated in the mitochondrial genomes of ancestral anaerobic kinetoplastids (Cavalier-Smith, 1997). Another model suggests that editing at the RNA level provides flexibility for sequence and functional adaptation, allowing for

increased genome mutation rates, genetic variation and selection (Landweber and Gilbert, 1993). RNA editing might also function as a regulatory mechanism to optimize energy metabolism in the different life stages of the kinetoplastid (Stuart et al., 1997). In African trypanosomes, different mitochondrial mRNAs are edited during the life cycle stage that employs oxidative phosphorylation as compared to stages in which glycolysis is utilized for energy production (Feagin et al., 1987). Though its evolutionary history remains unclear, kinetoplastid mRNA editing presents a remarkable example of atypical management of biological information.

1.1.2. Base Modification Editing

While uridine insertion/deletion editing has been observed only in kinetoplastids, base modification RNA editing occurs in all forms of life. Though many site-specific nucleotide modifications (e.g. methylation, acetylation, glycosylaton) occur in RNA (Rozenski et al., 1999), base modification RNA editing traditionally refers to the enzymatic deamination of an encoded base to create a substitute nucleotide. Polynucleotide cytidine deaminases convert cytidine to uridine (C-to-U, Figure 1.1), while polynucleotide adenosine deaminases convert adenosine to inosine (A-to-I, Figure 1.2). In both cases, the RNA phosphodiester backbone remains intact.

In the evolution of prokaryotes and eukaryotes, tRNA represents a common progenitor substrate for RNA editing. In bacteria and chloroplasts, $tRNA^{Arg2}$ contains an inosine nucleotide at position 34, the first (5') base of the anticodon sequence (Limbach et al., 1994; Sprinzl et al., 1996). Similarly, seven or eight tRNAs contain inosine at position 34 in eukaryotes. Inosine can base pair

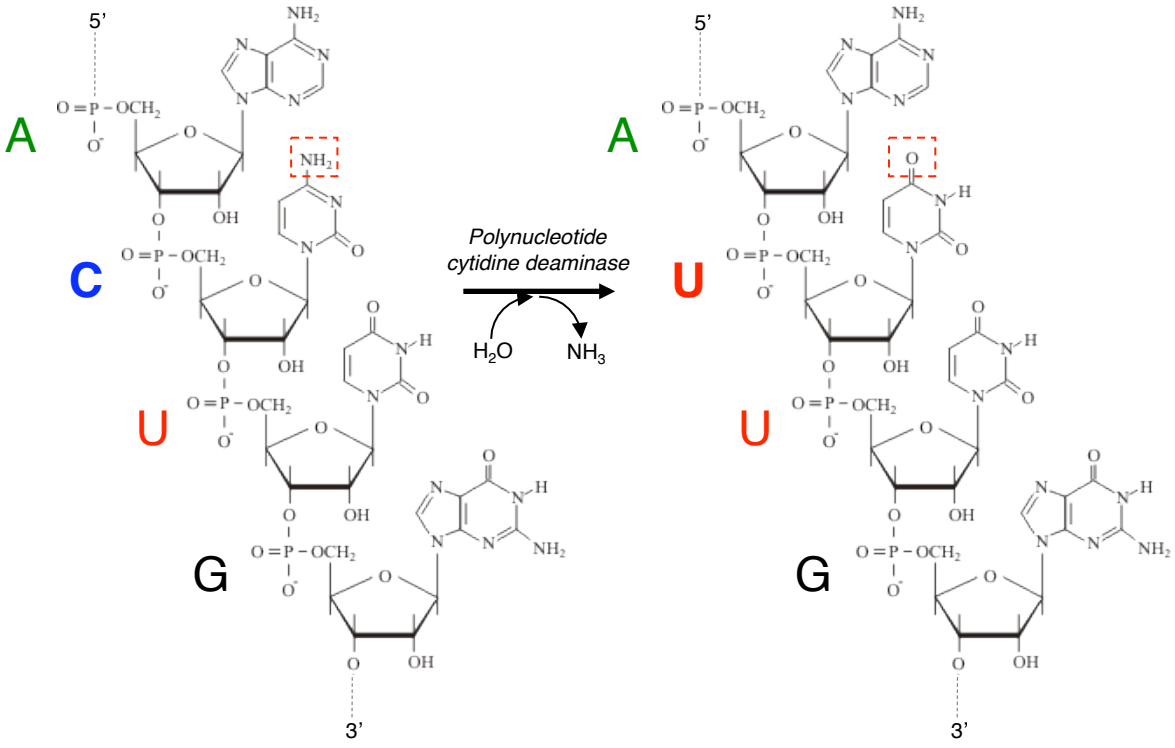


Figure 1.1. C-to-U editing by polynucleotide cytidine deaminases. The cytidine deaminase reaction, shown here for an RNA polynucleotide.

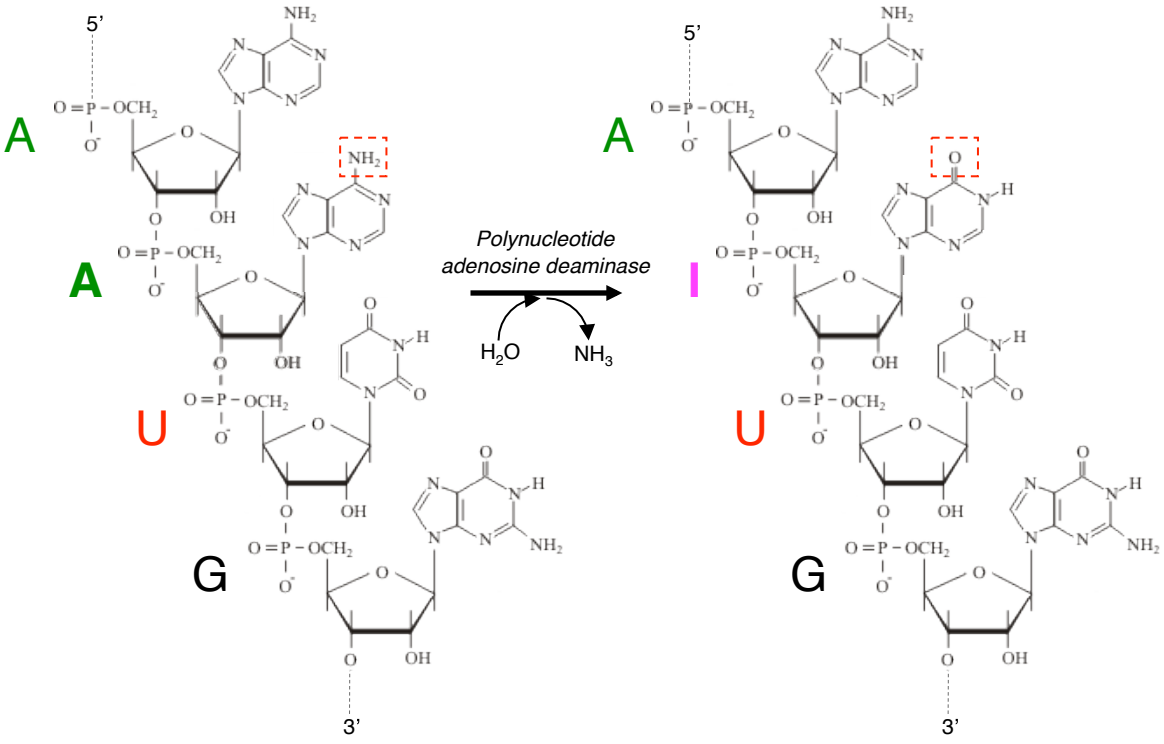


Figure 1.2. A-to-I editing by polynucleotide adenosine deaminases. The adenosine deaminase reaction, shown here for an RNA polynucleotide.

with three different RNA nucleotides: U, C, and A. This non-standard coupling property, along with G:U base pairing, forms the mechanistic basis for Crick's wobble hypothesis (Crick, 1966), in which the degeneracy of the genetic code is partially a result of single tRNA anticodons recognizing multiple, synonymous mRNA codons. The genes encoding these tRNAs all contain an adenosine at position 34; inosine is introduced by post-transcriptional RNA editing.

tRNA anticodon editing is carried out by adenosine deaminases acting on tRNA (ADATs). In bacteria, homodimeric TadA/ecADAT2 catalyzes the A-to-I editing at position 34 of $tRNA^{Arg2}$ (Wolf et al., 2002). The site-specificity of this reaction is determined primarily by the sequence of the anticodon loop substrate. *tadA* is essential for cell viability, which underscores the functional importance of wobble recognition in even a single tRNA. The TadA homolog in eukaryotes is a heterodimer (Tad2p/ADAT2 and Tad3p/ADAT3), which edits multiple tRNAs and exhibits broader specificity than its evolutionary predecessor (Gerber and Keller, 1999). ADAT1, a related eukaryotic enzyme, edits $tRNA^{Ala}$ at an adenosine nucleotide at position 37, immediately adjacent to the anticodon sequence (Gerber et al., 1998; Maas et al., 1999). The functional consequence of this modification is not clear, though it may be important for suppressing translational frameshift errors. Though adenosine to inosine changes constitute only a small fraction of the numerous base modifications present in tRNAs, they serve an essential and fundamental role, as evidenced in part by the broad evolutionary conservation of the ADAT editing machinery.

As perhaps the first evolutionary instance of polynucleotide RNA editing enzymes, the ADAT family provides mechanistic insight into the subsequent emergence of editors that act on mRNA and DNA substrates. Despite the

adenosine-to-inosine activity of the ADATs, their catalytic domains bear considerable similarity to those of the cytidine deaminases (Gerber and Keller, 1999). This characteristic zinc-dependent deaminase motif, [H/C]xEx$_{25\text{-}30}$PCxxC (where X is any amino acid), is also present in the prokaryotic mononucleotide cytidine deaminases (Betts et al., 1994; Ireton et al., 2002; Johansson et al., 2002; Xiang et al., 1997). An evolutionary shift in ADAT substrate specificity (i.e. from C to A) was probably the result of changes to the nucleotide binding properties of the enzyme active site. The sequence and structural similarities of adenosine and cytidine polynucleotide deaminases underscore the biochemical similarities of these two forms of base modification editing. Indeed, the trypanosomatid ADAT2/3 complex has the flexibility to perform both A to I editing in tRNA as well as C to U editing in DNA (Rubio et al., 2007). This permissive specificity provides functional evidence of an evolutionary relationship in which the cytidine deaminase motif-containing ADAT family diverged into the ADAR adenosine deaminases and AID/APOBEC polynucleotide cytidine deaminases (Conticello, 2008; Conticello et al., 2007).

1.2. Adenosine Deaminases Acting on RNA (ADARs)

The ADAR family of RNA editing enzymes emerged early in metazoan evolution, a result of ADATs gaining double-stranded RNA binding function (Jin et al., 2009). ADARs have been identified in most multicellular organisms examined, including worm (Tonkin et al., 2002), fly (Palladino et al., 2000b), fish (Slavov et al., 2000a; Slavov et al., 2000b), frog (Bass and Weintraub, 1987, 1988; Rebagliati and Melton, 1987), bird (Herbert et al., 1995), rodent (Melcher et al., 1996b; O'Connell et al., 1995) and human (Kim et al., 1994). The ADAR proteins contain a variable number of double-stranded RNA binding motifs (dsRBMs) followed by a C-terminal catalytic domain with similarities to the cytidine deaminase motif described above. ADARs bind to RNA duplex structures and catalyze the conversion of adenosine to inosine. A-to-I modification is the most common mRNA editing event in higher eukaryotes, and is important if not essential for animal life. Deletion of ADAR genes in mice results in embryonic lethality (ADAR1) (Hartner et al., 2004; Wang et al., 2004b) or severe neuropathology and early post-natal mortality (ADAR2) (Higuchi et al., 2000). *Drosophila melanogaster* mutants lacking ADAR activity display severe behavioral defects, seizures and paralysis (Palladino et al., 2000a). Such dramatic phenotypes underscore the significance of proper information management in the transcriptome and the impact of RNA editing on its regulation.

Unlike their ADAT evolutionary precursors, ADARs act on a diverse array of cellular transcripts. Three ADAR proteins have been identified in mammals; ADAR1 and ADAR2 both demonstrate broad specificity for editing targets (Riedmann et al., 2008), while ADAR3 does not exhibit detectable catalytic activity (Chen et al., 2000; Melcher et al., 1996a). ADAR editing targets

include mRNA coding sequences, mRNA untranslated regions, splice sites, introns, transcribed *Alu*/SINE elements, pre-miRNA transcripts and viral RNA. A-to-I modification in these different RNA substrates can result in various functional consequences.

1.2.1. mRNA Coding Sequence Targets of ADAR Editing

ADAR targets in mRNA coding sequences provide the most direct examples of editing impact on biological outcomes. Inosine is read as guanosine by the translational machinery (Basilio et al., 1962), so A-to-I editing effectively induces A to G sequence alterations in protein coding information. Editing events that result in non-synonymous codon changes contribute to considerable proteome diversity in several tissues, particularly brain. For example, ADAR2 edits the mRNA encoding the glutamate-gated ion channel AMPA receptor subunit GluR-B at a specific adenosine termed the Q/R site (Sommer et al., 1991). Upon translation, inosine at this position results in an ion channel protein product with notable functional differences relative to the non-edited isoform, including decreased calcium permeability and altered channel dynamics (Lomeli et al., 1994). The severe neuropathologic phenotype of *adar2*$^{-/-}$ mice can be rescued by introducing a single point mutation at the Q/R site within the GluR-B gene, demonstrating the physiological significance of this editing event (Higuchi et al., 2000). Though an arginine residue at the GluR-B Q/R position is critical in the adult brain, editing provides a functional flexibility that may be important in brain development and plasticity. Such flexibility is apparent in many other ADAR-targeted gluR subunit mRNAs, for which edited and non-edited isoforms

coexist and can be developmentally regulated (Bernard and Khrestchatisky, 1994; Lomeli et al., 1994).

The 5-HT_{2C}R serotonin receptor mRNA is another neurotransmitter transcript that undergoes coding sequence changes as a consequence of ADAR editing. 5 different A-to-I editing sites are present in the 5-HT_{2C}R mRNA, each of which results in a single amino acid substitution (Burns et al., 1997; Niswender et al., 1998). The protein sequence changes alter the affinity of the serotonin receptor for G protein coupling, which in turn affects ligand binding affinities and downstream signaling. Combinatorial editing can lead to many different 5-HT_{2C}R protein isoforms (Burns et al., 1997; Niswender et al., 1999), which exist at different levels in different parts of the brain (Burns et al., 1997). The variety obtained from a single primary transcript demonstrates the potential impact of RNA editing on proteome diversification, a particularly important feature in an exceedingly complex tissue such as the brain.

1.2.2. Untranslated RNA Sequence Targets of ADAR Editing

In addition to neurotransmitter receptors, ADAR coding sequence editing has been observed in several other cellular mRNAs for proteins of various function, including the actin-binding protein filamin A, bladder cancer associated protein (blcap) and insulin-like growth factor binding protein 7 (Levanon et al., 2005; Riedmann et al., 2008). However, recent bioinformatic (Athanasiadis et al., 2004; Blow et al., 2004; Kim et al., 2004; Levanon et al., 2004) and ultra-high throughput sequencing analyses (Li et al., 2009b) have shown that most A-to-I mRNA editing occurs in untranslated transcript regions and non-coding RNAs. In particular, *Alu* elements in transcript 3′ UTRs and introns are

the most frequent targets of A-to-I editing in a variety of tissues. *Alu* sequences are short repetitive retrotransposable elements interspersed throughout primate genomes (Cordaux and Batzer, 2009). *Alu* sequences do not code for protein, but are often transcribed within the untranslated regions and introns of mRNAs. More than 10,000 A-to-I editing sites have been identified in *Alu* sequences (Levanon et al., 2004), though many of these occur within introns that may be spliced out and therefore absent from mature transcripts.

The functional consequences of A-to-I changes in *Alu* and other untranslated sequences are somewhat unclear, but recent work suggests that editing may play a role in gene regulation. Editing in transcript 3′ UTRs correlates with a reduction in protein expression (Chen et al., 2008). The specific mechanism for decreasing expression remains undefined, but may result from the nuclear retention of those transcripts with inosine-containing 3′ UTRs (Chen and Carmichael, 2008). Other possibilities include sequence-dependent changes in RNA stability, structure and/or recognition by regulatory RNA binding proteins. Examples of A-to-I editing that generate miRNA seed target sequences in transcript 3′ UTRs have also been described (Borchert et al., 2009). Regardless of mechanism, the observation that post-transcriptional alterations in untranslated regions can affect mRNA processing and/or translation serves as a prime example of the functional significance of non-coding sequences. Though such sequences do not directly code for polypeptide assembly, they can contain information relevant to transcript and protein regulation.

1.2.3. Viral RNA targets of ADAR Editing

A-to-I RNA editing is not limited to cellular transcripts; ADARs act on viral RNA as well. ADAR1 expression is inducible by Type I interferon, suggesting a likely function in the host antiviral response. However, in different contexts, it seems that ADAR targeting of viral RNA can benefit the host cell and/or the infecting virus. A to I editing has been observed in diverse viral RNAs, including influenza virus (Tenoever et al., 2007), parainfluenza virus (Murphy et al., 1991), lymphocytic choriomeningitis virus (LCMV) (Zahn et al., 2007), vesicular stomatitis virus (O'Hara et al., 1984), measles virus (Baczko et al., 1993; Cattaneo et al., 1988), polyomavirus (Kumar and Carmichael, 1997), and hepatitis D virus (HDV) (Luo et al., 1990). Despite an early recognition of A to I biased hyperediting in viral transcripts during persistent and lytic infections (Cattaneo, 1994), the function and consequences of many of these editing events is still under investigation. A clear example of direct ADAR antiviral editing has been observed in LCMV RNA transcripts, in which large numbers of A-to-I mutations disrupt glycoprotein coding sequences and impair viral infectivity (Zahn et al., 2007). A-to-I editing may also impair RNA viral infection though an inosine-specific RNAse (Scadden and Smith, 1997, 2001). Additional evidence suggests that inosine-containing viral RNA might recruit specific RNA-binding proteins that can suppress replication and/or translation (Zhang and Carmichael, 2001).

ADAR editing of viral RNA is not always inhibitory to the virus. In apparent contradiction to the anti-viral functions described above, ADAR proteins can also promote viral replication and production. In HIV infection, ADAR1 enhances viral protein expression, replication, and infectivity through

both editing-dependent and editing-independent mechanisms (Doria et al., 2009). Specific targets of A-to-I editing in HIV RNA have been identified in the 5′ UTR shared by all HIV transcripts (Doria et al., 2009), as well as in an *env* gene protein coding sequence (Phuphuakrat et al., 2008). In HDV infection, ADAR1 editing is a necessary process in the virus life cycle. Though it co-opts hepatitis B virus (HBV) surface antigen for capsid formation, HDV also encodes its own surface antigen (HDVAg). HDVAg occurs in short (HDVAg-S) and long (HDVAg-L) isoforms, essential for viral replication (Kuo et al., 1989) and packaging (Chang et al., 1991; Ryu et al., 1992), respectively. These two isoforms are regulated by ADAR1 editing, which targets a specific "amber/W site" in the HDV antigenome to convert a UAG stop codon (translating to HDVAg-S) to a UIG tryptophan codon (translating to HDVAg-L) (Casey and Gerin, 1995; Luo et al., 1990; Polson et al., 1996). This editing event serves as a regulatory switch that shifts the HDV infection cycle from viral genome replication to virion assembly and packaging.

1.2.4. Sequence and Structural Characteristics of ADAR Editing Targets

With diverse targets in coding exons, untranslated regions, introns and viral RNAs, ADAR enzymes appear to have a broad specificity for many different sequence types. However, all ADAR targets share several characteristic sequence and structural features that determine enzyme binding and editing specificity. ADAR editing sites always occur in double-stranded RNA (dsRNA) structures (Hundley and Bass, 2010). In fact, prior to recognition of their editing activity, the ADARs were first identified as dsRNA-unwinding enzymes (Bass and Weintraub, 1987; Rebagliati and Melton, 1987). RNA duplex binding is

dictated by the dsRBMs present in all ADAR enzymes (Stephens et al., 2004; Xu et al., 2006). However, not all dsRNA structures support ADAR editing. To qualify as an ADAR substrate, an RNA duplex must be of sufficient length to allow binding of all dsRBMs (Hundley and Bass, 2010). In the case of ADAR2, the duplex must also be long enough to support the binding of a second ADAR2 monomer, thereby establishing the dimeric complex required for editing (Chilibeck et al., 2006; Cho et al., 2003; Gallo et al., 2003; Jaikaran et al., 2002). Within a perfectly paired RNA duplex of sufficient length, ADAR editing is non-specific (Nishikura et al., 1991; Polson and Bass, 1994); multiple adenosines will be deaminated until the double-stranded structure is unwound due to I:U mismatches (Bass and Weintraub, 1988; Wagner et al., 1989). However, site-specific editing is defined by structural features such as bulges, mismatched base pairs, and internal loops (Dawson et al., 2004; Lehmann and Bass, 1999; Ohman et al., 2000), as well as the nucleotides immediately adjacent to the edited adenosine. With respect to an editing site, both ADAR1 and ADAR2 exhibit a similar preference for the 5′ neighboring base (A = U > C > G) (Lehmann and Bass, 2000; Polson and Bass, 1994). ADAR2 also has a preference for the 3′ neighboring base (G = U > C =A) (Lehmann and Bass, 2000). These flanking base preferences have been consistently reinforced as additional ADAR editing targets have been identified.

Targets of ADAR editing such as those discussed above fulfill these substrate criteria through local or distant, *cis* or *trans* RNA sequences. For ADAR editing within coding regions, complementary sequence in an adjacent intron often provides a source of base pairing for duplex formation (Herb et al., 1996; Higuchi et al., 1993). Transcript 3′ UTRs frequently contain considerable

secondary structure, providing appropriate dsRNA for ADAR editing. In addition, duplicate, inverted *Alu* elements common in non-coding RNA sequences can hybridize into long dsRNA duplexes that support editing (Athanasiadis et al., 2004; Kawahara and Nishikura, 2006; Kim et al., 2004; Levanon et al., 2004). Analogous structures are found in the double-stranded replication intermediates of many RNA viruses, as well as pre-miRNA transcripts, some of which have recently been described as ADAR substrates (Blow et al., 2006; Kawahara et al., 2007; Luciano et al., 2004; Yang et al., 2006b). Subtle differences within these common structures and sequences can influence editing efficiency, resulting in varied proportions of edited and unedited transcripts for a given target site. Taken together, the specific yet flexible substrate restrictions and variable frequency of A-to-I editing allow for widespread and precise sequence adjustments to the transcriptome by ADAR enzymes.

1.3. Polynucleotide Cytidine Deaminases: The AID/APOBEC enzymes

The AID/APOBEC enzymes also perform base modification editing on polynucleotide substrates. However, unlike the ADARs, AID/APOBEC family members act on cytidine residues, which they deaminate to form uridine. This editing activity is not limited to RNA; most AID/APOBEC enzymes deaminate single-stranded DNA as their primary substrate. Understanding APOBEC RNA editing is best achieved with a comprehensive background of the entire enzyme family (including those members that act on DNA), which is presented here.

Though the ADAR gene family appears to be present in all metazoans (Jin et al., 2009), the evolutionarily younger AID/APOBEC polynucleotide cytidine deaminases emerged later, in the vertebrate lineage (Conticello et al., 2005). In mammals, this gene family includes Activation-induced cytidine deaminase (AID), Apolipoprotein B mRNA Editing Catalytic polypeptide 1 (APOBEC1), APOBEC2, APOBEC3 and the computationally predicted APOBEC4. Additionally, in primates, the APOBEC3 gene has undergone considerable expansion; genes in this subfamily are designated alphabetically APOBEC3A-3H. AID and APOBEC2 are the oldest members of the gene family, with the appearance of APOBEC1 and APOBEC3 occurring significantly later in mammalian evolution (Conticello et al., 2005). In primates, members of the AID/APOBEC gene family have undergone rapid evolution and notably strong positive selection (Sawyer et al., 2004), likely due to functions for many of these enzymes in host defense.

All AID/APOBEC proteins contain at least one characteristic zinc-dependent deaminase domain (Jarmuz et al., 2002; Mian et al., 1998; Wedekind et al., 2003). The active site includes three critical residues (2 cysteines, 1 histidine)

that coordinate zinc to activate a water molecule for the hydrolytic deamination of cytidine. An additional conserved glutamic acid residue participates in the transfer of a water-derived proton to the leaving imino group of the cytidine ring (Smith, 2009). The catalytic protein motif and mechanism are similar to those of bacterial cytidine deaminases that act on mononucleotide substrates (Navaratnam et al., 1995). Most AID/APOBEC proteins contain a single zinc-dependent deaminase motif, though several of the APOBEC3 enzymes (murine APOBEC3, primate APOBEC3B, APOBEC3DE, APOBEC3F, APOBEC3G) contain two in distinct domains. In both APOBEC3F and APOBEC3G, only one of the two predicted active sites demonstrates catalytic activity (Haché et al., 2005).

The AID/APOBEC cytidine deaminases contribute to diverse biological processes, many of which are involved in host defense functions. AID introduces DNA mutations at the immunoglobulin (Ig) locus, which is critical in establishing the genetic diversity required for an effective humoral immune response. APOBEC3 enzymes edit retroviral target sequences and thereby provide a potent innate defense that can mutate and cripple viral genomes. In addition, APOBEC1 edits the apoB mRNA transcript, mediating the tissue-specific regulation of apolipoprotein isoforms important in dietary lipid absorption. Though each of these processes shares a common feature of cytidine to uridine conversion, the physiological contexts and outcomes vary widely.

1.3.1. AID - Editing for Antibody Diversification

AID is a central mediator of the adaptive immune response, driving antibody diversification in response to antigenic challenge through two distinct processes: somatic hypermutation (SHM) and class switch recombination (CSR).

In SHM (Figure 1.3A), point mutations are introduced within the V(D)J regions of rearranged Ig loci, thereby altering the antigen binding properties of the encoded Ig. SHM occurs in germinal center B cells and those B cells that acquire improved antigen binding affinity are positively selected for survival. During this "affinity maturation," the Ig variable regions accumulate additional mutations and the activated B cells are thereby optimized for a highly specific and potent adaptive immune response. As these modifications are fixed in somatic B cell genomes, they are maintained in clonal expansion and subsequent memory responses. In CSR (Figure 1.3B), a region specific recombination reaction replaces the primary constant region (Cμ) with a downstream constant region (Cγ , Cε , or Cα). Constant regions code for the Fc portions of antibody molecules; recombination between these regions changes the isotype and endows the antibody protein with different effector properties. This genomic and functional flexibility allows for an antigen-appropriate response in many different infection contexts. Both of these processes are initiated by AID cytidine deamination within the Ig locus.

Initially identified in a subtractive hybridization screen comparing resting B cells and B cells stimulated to undergo CSR (Muramatsu et al., 1999), AID acts on cytidine residues in single-stranded DNA (Chaudhuri et al., 2003; Dickerson et al., 2003; Pham et al., 2003). In germinal center B cells, transcription at the Ig locus separates the genomic DNA duplex, thereby generating a single-stranded substrate for AID deamination. In SHM, AID activity introduces U:G mismatches within rearranged Ig V(D)J gene segments. The cellular DNA repair machinery recognizes mismatches and the error-prone resolution of these

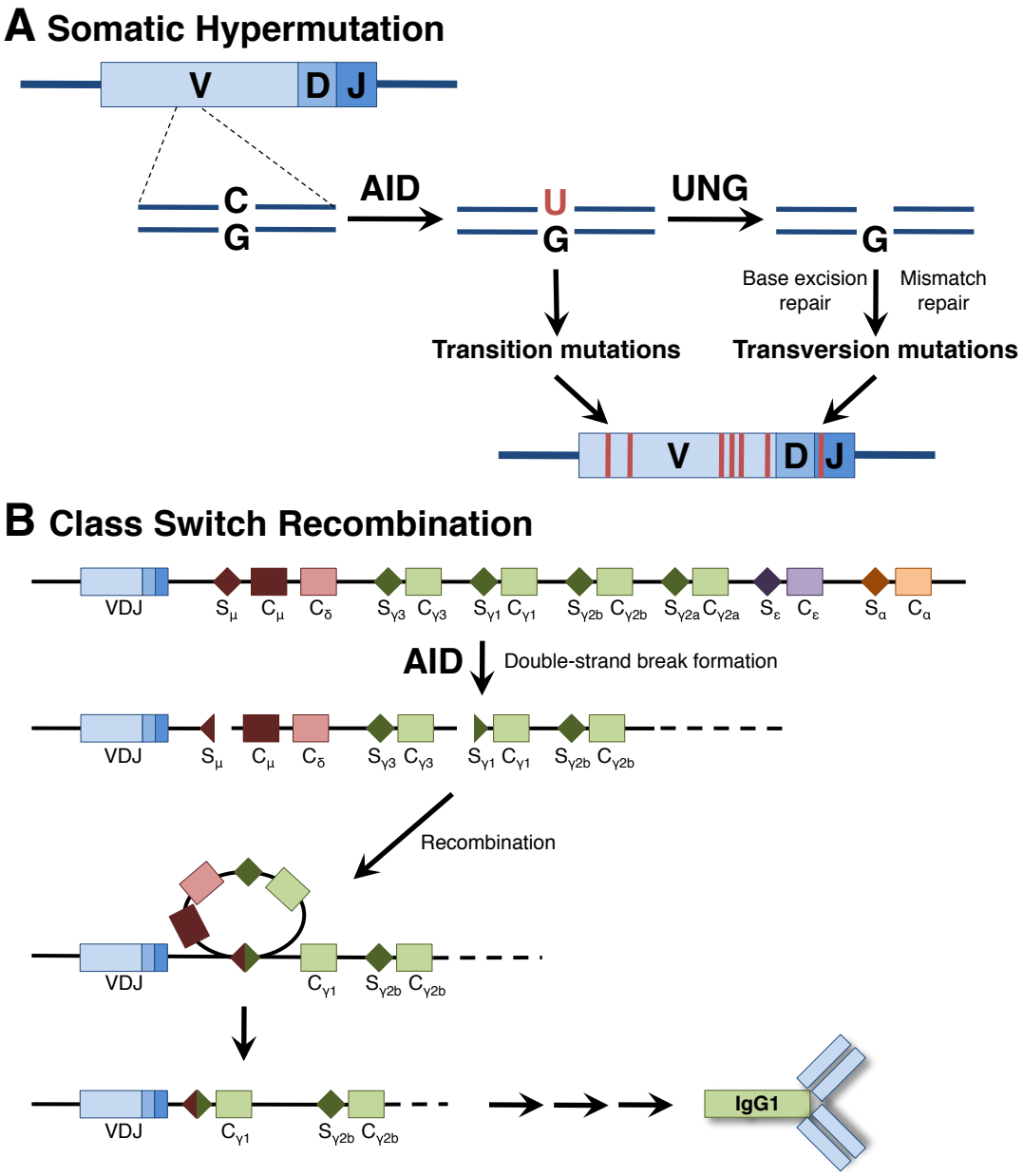


Figure 1.3. AID drives antibody diversity through two distinct mechanisms. (A) In SHM, AID deaminates cytidines within the variable region of Ig loci. Genomic deoxyuridine residues are then resolved by two pathways. Uridine is read as thymidine by replication machinery, leading to transition mutations. Alternatively, UNG excises the edited base, which is followed by abasic site processing via base excision repair and mismatch repair enzymes, leading to transversion mutations. (B) In CSR, AID deaminates cytidines within Ig switch (S) regions, leading to double strand breaks. Recombination replaces the primary constant switch region (C_μ) with one of several downstream constant regions (C_γ, C_ε, or C_α), altering the effector properties of the encoded antibody. In this representation, the primary C_μ region is replaced with a $C_{\gamma 1}$ region, thereby causing a switch from the IgM to the $IgG_{\gamma 1}$ isotype.

deoxyuridines leads to the transition and transversion point mutations of B cell affinity maturation. In CSR, AID deaminates cytidines within repetitive sequences between the variable and constant regions of Ig gene segments. This editing event (and corresponding DNA breaks) initiates the region-specific recombination reaction that results in Ig constant region isotype switch.

AID is necessary for both SHM and CSR; these processes are completely absent in *aicda*$^{-/-}$ (AID knockout) mice (Muramatsu et al., 2000). This functional dependence is also observed in humans with Hyper-IgM syndrome, an autosomal recessive immunodeficiency characterized by the absence of CSR with corresponding elevated serum levels of IgM, the lack of SHM, lymph node hyperplasia, and increased susceptibility to infection by encapsulated bacteria. The Type II variant of Hyper-IgM syndrome is caused by mutations in the AID gene (Revy et al., 2000). Further evidence for AID's central role in SHM and CSR is provided by experiments demonstrating that exogenous AID expression is necessary and sufficient to induce both processes in hybridoma cell lines (Martin et al., 2002) and fibroblasts (Okazaki et al., 2002).

Deficiencies in DNA repair components also affect SHM and CSR. As a result of cytidine deamination, AID generates U:G mismatches, which are typically resolved by Uracil DNA Glycosylase (UNG), the primary effector of uracil removal in base excision repair. While UNG-deficient animals acquire Ig locus mutations at rates comparable to wild-type controls , the spectra of mutations are dramatically different, with a strong bias towards G-to-A and C-to-T events (Rada et al., 2002). Ig class-switching is also dramatically reduced. Once again, a human analogue is evident in Type V Hyper-IgM syndrome, which is associated with mutations UNG gene mutations (Imai et al., 2003; Lee et

al., 2005). Several other DNA repair factors, including MSH2 (Ehrenstein and Neuberger, 1999; Rada et al., 1998), MSH6 (Martomo et al., 2004; Martomo et al., 2005), DNA polymerase η and exonuclease 1 have also been implicated in SHM and/or CSR, further underscoring the shared contributions of AID cytidine deamination and DNA repair processes in shaping the antibody repertoire.

As a mutator of genomic DNA, AID must be tightly regulated in germinal center B cells. Though AID can be observed throughout the cell (Ito et al., 2004), it is sequestered in the cytoplasm away from genomic DNA by active nuclear export (Brar et al., 2004) and cytoplasmic retention mechanisms (Patenaude et al., 2009). Conformational changes in an atypical nuclear localization signal allow for the active nuclear import of AID (Patenaude et al., 2009). Once inside the nucleus, the mechanism by which AID specifically targets Ig loci remains unclear. Several associated *cis* chromosomal elements can contribute to AID targeting, such as the IgV promoter (Yang et al., 2006a) and Ig heavy chain 3' enhancer sequences (Dunnick et al., 2009), but are not sufficient to explain AID specificity (Yang and Schatz, 2007). Alternatively or additionally, recent observations of AID DNA editing at numerous genome locations suggest that AID may lack stringent targeting specificity, with an exclusion of high-fidelity DNA repair machinery only at the Ig locus allowing for local accumulation of mutations (Liu et al., 2008). Once associated with a ssDNA substrate, AID exhibits local sequence preferences for its deaminase activity. AID preferentially deaminates cytidine residues contained in WRC mutation "hot spot" motifs, while avoiding activity on cytidines in SYC "cold spot" motifs (Beale et al., 2004; Pham et al., 2003). Similar preferences for neighboring nucleotides have also been described for related cytidine deaminases such as APOBEC3G and

APOBEC3F (Beale et al., 2004; Harris et al., 2003; Mangeat et al., 2003; Zennou and Bieniasz, 2006; Zhang et al., 2003).

Despite benefits in host defense, the risk of off-target genomic DNA editing is significant. In the context of CSR, it is thought that AID deamination (together with UNG activity) generates coupled DNA breaks at the Ig locus. Most of these are appropriately repaired, leading to deletional recombination. However, occasionally the breaks may be erroneously resolved, giving rise to chromosomal translocations. The partner loci for such rearrangements can be either genomic sites that can serve as infrequent off-site targets for AID (including many proto-oncogenes (Liu et al., 2008; Robbiani et al., 2008)), or hypersensitive sites susceptible due to structure (e.g. cruciform DNA at the bcl-2 locus (Raghavan et al., 2004)) or function (e.g. enhancer elements). Incorrect break resolution can be caused by deficiencies of DNA repair factors (Ramiro et al., 2006) or lack of Ig locus CSR targeting elements (Wakae et al., 2006). Oncogenic events represent an infrequent but deleterious outcome of dysregulated and/or mistargeted DNA editing. However, in the case of AID, the benefit of antibody diversification to host defense generally outweighs these consequences of off-target mutation.

1.3.2. APOBEC3 Enzymes - Innate Defense Against Retroviruses

Initially identified by bioinformatic sequence analysis (Jarmuz et al., 2002), the APOBEC3 proteins did not have an obvious function upon first examination. Though recognized as polynucleotide cytidine deaminases with likely RNA and/or DNA editing activity, their targets were unknown. However, initially

unrelated work in HIV biology soon revealed the potent effects of the APOBEC3 enzymes on retroviral infection.

As a complex lentivirus, HIV is built from a genome containing the fundamental retroviral genes *gag*, *pol* and *env*, as well as several accessory genes that enhance viral infectivity (Malim and Emerman, 2008). Most of these accessory genes were successfully characterized relatively soon after their identification. However, a mechanistic function for virion infectivity factor (Vif) remained elusive. It was observed that Δ*vif* HIV infection is essentially unhindered in certain cell lines (so-called "permissive cells") yet dramatically diminished in other cell lines ("non-permissive cells") that otherwise support productive infection with wild-type HIV (Gabuzda et al., 1992);von Schwedler, 1993, p05754}. Heterokaryon studies suggested that the non-permissive cells expressed an endogenous viral restriction factor that was apparently obstructed by Vif (Madani and Kabat, 1998; Simon et al., 1998). A subtractive cDNA screen performed by Sheehy and colleagues identified this cellular restriction factor as CEM15 (Sheehy et al., 2002), also known as APOBEC3G.

APOBEC3G's place in the cytidine deaminase family provided clues as to how it might function to restrict Δ*vif* HIV infection. A torrent of work from various groups demonstrated that APOBEC3G restricts HIV by inducing hypermutation in newly reverse-transcribed viral cDNA (Figure 1.4) (Harris et al., 2003; Lecossier et al., 2003). In non-permissive cells infected with Δ*vif* HIV, APOBEC3G associates with viral RNA (Khan et al., 2005; Svarovskaia et al., 2004) and/or the nucleocapsid domain of Gag (Schäfer et al., 2004), which package the enzyme into newly assembled virions. Nascent viral particles then carry and deliver APOBEC3G to newly infected cells, in which the cytidine deaminase

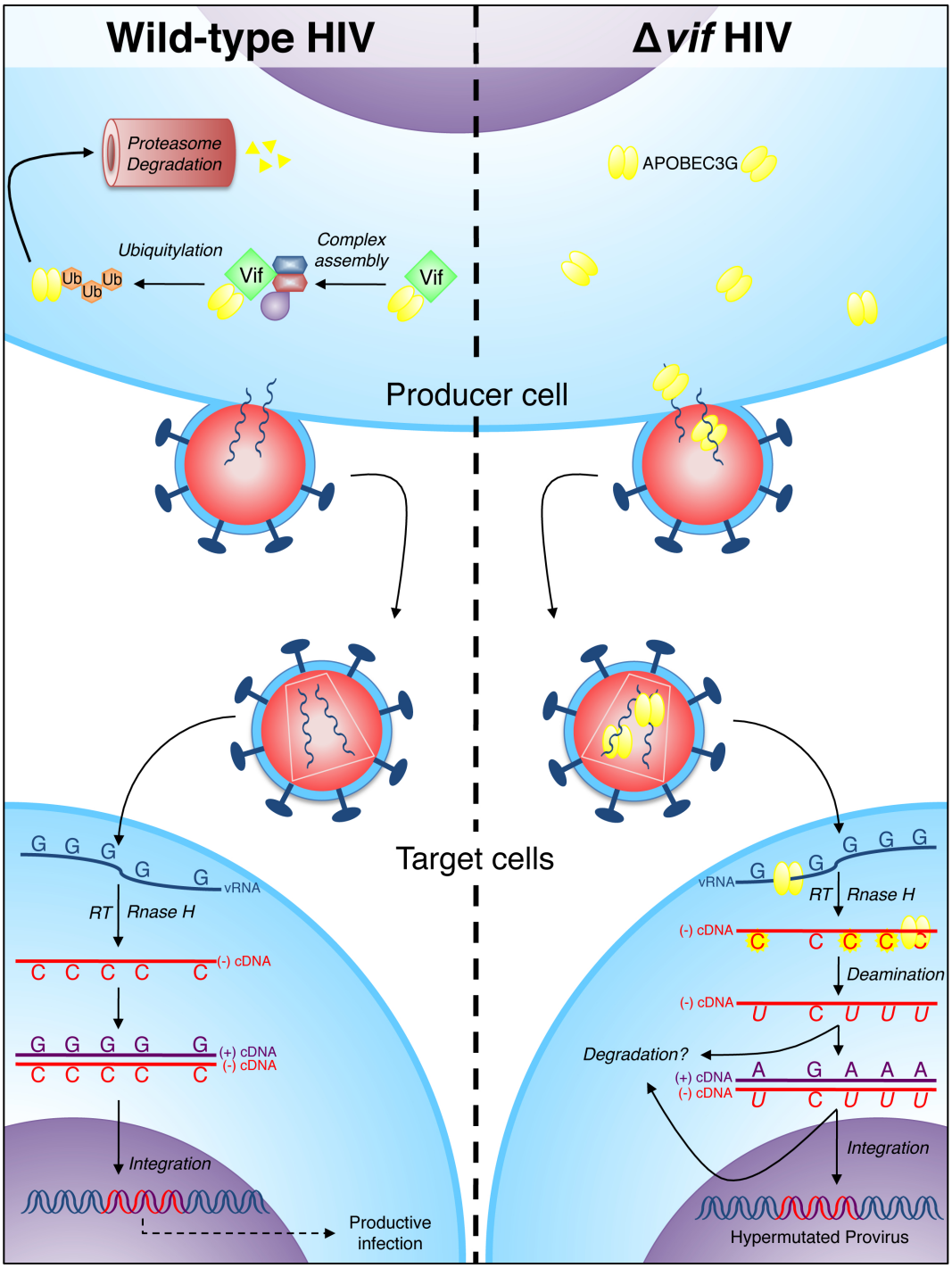


Figure 1.4. APOBEC3G restricts HIV infection. In the context of wild-type HIV infection, HIV Vif binds cellular APOBEC3G and recruits a cullin5-elongin B/C-Rbx ubiquitin ligase complex. APOBEC3G is polyubiquitylated and targeted for degradation in the proteasome. In the absence of HIV Vif, intact APOBEC3G packages with nascent virus particles and is delivered to newly infected cells. Upon reverse transcription, APOBEC3G edits newly synthesized (-) strand viral cDNA. Edited retrotranscripts are degraded or integrated as hypermutated proviruses.

remains associated with the retroviral replication machinery. Upon reverse transcription and following activation by retroviral RNAseH activity (Soros et al., 2007), APOBEC3G actively deaminates along the retroviral (–)-strand cDNA, introducing C-to-U changes through the viral retrogenome. Most hypermutated (–)-strand cDNAs do not proceed to second-strand synthesis, though the mechanism for their degradation remains unclear. Those retrotranscripts that do progress to integration carry a heavy mutation burden that drastically compromises the information content of the retroviral genome and prevents the production of functional virus.

Both $CD4^+$ T cells and macrophages, the principle targets of HIV infection *in vivo*, express high levels of APOBEC3G (Chiu et al., 2005). HIV establishes productive infections in these cell types through the action of Vif, which excludes APOBEC3G from packaging virions primarily by mediating its degradation (Conticello et al., 2003; Marin et al., 2003; Mehle et al., 2004; Sheehy et al., 2003; Stopak et al., 2003). In the infected cell, Vif binds APOBEG3G and recruits a cullin5-elongin B/C-Rbx ubiquitin ligase complex (Yu et al., 2003) (Kobayashi et al., 2005). APOBEC3G is thereby targeted for ubiquitylation and subsequent degradation by the proteasome. This process allows for the production of virions unfettered by packaged APOBEC3G and corresponding hypermutation. The Vif/APOBEC3G interaction is precise and species-specific; HIV Vif inhibits APOBEC3G from human but not African Green Monkey (AGM), while SIV-AGM Vif inhibits AGM APOBEC3G, but not human (Mariani et al., 2003). This specificity maps to a single amino acid at position 128 in APOBEC3G (Asp in human, Lys in AGM) (Bogerd et al., 2004; Mangeat et al., 2004; Schröfelbauer et al., 2006). Such exacting requirements for association offer a perspective on the

powerful selective pressures exhibited by this host-pathogen relationship. In fact, phylogenetic analysis demonstrates that the APOBEC3 subfamily has been under markedly strong positive selection throughout primate evolution (Sawyer et al., 2004; Zhang and Webb, 2004). It appears that the sole function of Vif is to act as a viral countermeasure against APOBEC3G and related family members. The evolutionary pressure to devote such a significant portion of the remarkably efficient HIV genome to this purpose illustrates the considerable impact this gene family can have on host defense.

The antiviral activity of the polynucleotide cytidine deaminases is not limited to APOBEC3G, nor is it directed solely against HIV. Aside from APOBEC3G, additional family members APOBEC3B and APOBEC3F can also restrict HIV replication (reviewed in Rosenberg and Papavasiliou, 2007). Interestingly, APOBEC3B cannot be suppressed by Vif (Bishop et al., 2004; Doehle et al., 2005). However, this may not be relevant *in vivo*, as only APOBEC3G and APOBEC3F are expressed by $CD4^+$ T cells and macrophages. Indeed, viral sequence analysis of HIV patient isolates reveals mutation patterns consistent with APOBEC3G and APOBEC3F activity (Simon et al., 2005).

Apart from HIV, it seems that most retroviruses are somewhat susceptible to members of the APOBEC3 subfamily; Murine leukemia virus, Equine infectious anemia virus, foamy retroviruses and even HBV can be restricted by different APOBEC3 deaminases (reviewed in Rosenberg and Papavasiliou, 2007). Though not a true retrovirus, HBV replicates its partially double-stranded DNA genome by a reverse transcription mechanism (Ganem, 1996), which represents a likely target for these deaminases. However, reverse transcription may not be a requirement for susceptibility to all APOBEC3 subfamily members; APOBEC3A

can dramatically inhibit replication of adeno-associated virus (AAV), a small single-stranded DNA parvovirus that replicates via the host cell polymerase machinery (Chen et al., 2006). It remains to be seen if AAV represents a unique exception or if APOBEC3 subfamily members exhibit broad antiviral activity against diverse non-retroviruses.

The suppressive activity of the APOBEC3 proteins is not limited to exogenous retroviruses; several family members have also been demonstrated to restrict a variety of endogenous retroelements. Such host-encoded retroelements constitute significant portions of mammalian genomes and include the long terminal repeat (LTR)-containing endogenous retroviruses (ERVs), as well as non-LTR sequences such as long interspersed nuclear element -1(LINE-1). Many of the APOBEC3 deaminases can suppress retrotransposition of ERVs (Esnault et al., 2005) and LINE-1 elements (Muckenfuss et al., 2006). Though it remains unclear if the precise mechanism by which APOBEC3 proteins inhibit endogenous retroelements is identical to that employed in exogenous viral restriction, both appear to be significant targets for this defense system.

As the existence of functionally active ERVs in humans remains uncertain, the physiological relevance of APOBEC3 subfamily activity against these host-encoded sequences *in vivo* is not immediately apparent. Perhaps retroelement inhibition represents an "original" function of the APOBEC3 proteins; activity against exogenous lentiviruses may have been co-opted later in the evolutionary history of this gene family. Indeed, despite dramatic rates of positive selection throughout primate evolution, the apparent selective pressures on the APOBEC3 genes predate the emergence of modern primate lentiviruses by millions of years (Sawyer et al., 2004; Zhang and Webb, 2004). In addition, the evolutionary

expansion of the APOBEC3 family correlates with a dramatic decrease of endogenous retroelement activity in primate genomes as compared to rodents (Waterston et al., 2002). Thus, the APOBEC3 proteins may represent an ancient defense system for protecting genome integrity, be it from actively mobile endogenous retroelements or the primitive retroviruses from which they evolved.

1.3.3. APOBEC2 and APOBEC4: Orphan Cytidine Deaminases

Along with AID, APOBEC2 is one of the oldest members of the polynucleotide cytidine deaminase gene family (Conticello et al., 2005). However, despite considerable sequence homology to other well-characterized AID/APOBEC proteins, the molecular and physiological functions of APOBEC2 remain largely unknown. Though initially predicted to act as a DNA/RNA editor (Liao et al., 1999), APOBEC2 does not deaminate polynucleotide substrates *in vitro* (Liao et al., 1999; Mikl et al., 2005). Additional biochemical characterizations have generated conflicting data, with some experiments demonstrating free mononucleotide deaminase activity (Anant et al., 2001b; Liao et al., 1999) and others suggesting a catalytically inert protein (Mikl et al., 2005). The APOBEC2 crystal structure revealed a homotetrameric polypeptide complex with active sites accessible to DNA and RNA (Prochnow et al., 2007), but while useful for modeling the structures of other AID/APOBEC family members, it has not elucidated APOBEC2 function. APOBEC2 is present exclusively in skeletal and cardiac muscle, where it is expressed at high levels (Liao et al., 1999). Upon initial examination, *apobec2*$^{-/-}$ mice were reported to be apparently normal and without phenotypic defects (Mikl et al., 2005). However, more thorough

investigation has recently revealed that APOBEC2 functions in muscle differentiation and maintenance (Sato et al., 2009). APOBEC2 is present at higher levels in slow-twitch (high aerobic potential) muscle fibers and *apobec2*$^{-/-}$ mice have dramatically higher slow-twitch:fast-twitch fiber ratios than wild-type controls. The molecular mechanism for these physiological differences remain unclear; if APOBEC2 edits RNA or DNA substrates, the identification of its target sequences should provide considerable insight as to its role in muscle development and function.

APOBEC4 is another "orphan" member of the polynucleotide cytidine deaminase gene family. Identified by computational sequence analysis, the APOBEC4 gene contains a characteristic zinc-dependent deaminase domain and bears significant homology to other family members (Rogozin et al., 2005). However, it's catalytic potential, substrate specificity (DNA/RNA) and possible sequence preferences remain unknown. APOBEC4 is predominantly expressed in testis, where its physiological function has not been identified. As with APOBEC2, the identification of editing targets for APOBEC4 in testis should greatly aid its biological characterization.

1.3.4. APOBEC1 – An RNA Editing Cytidine Deaminase

APOBEC1 was the first polynucleotide cytidine deaminase discovered in mammals, and is named for its editing substrate, apolipoprotein B (apoB) mRNA (Figure 1.5). The apoB protein product exists as two related forms, both important in lipid metabolism (Chan, 1992). The full-length isoform, apoB-100, is produced by the liver and forms the principle lipoprotein component of LDL particles, which transport endogenously synthesized triglycerides in the blood.

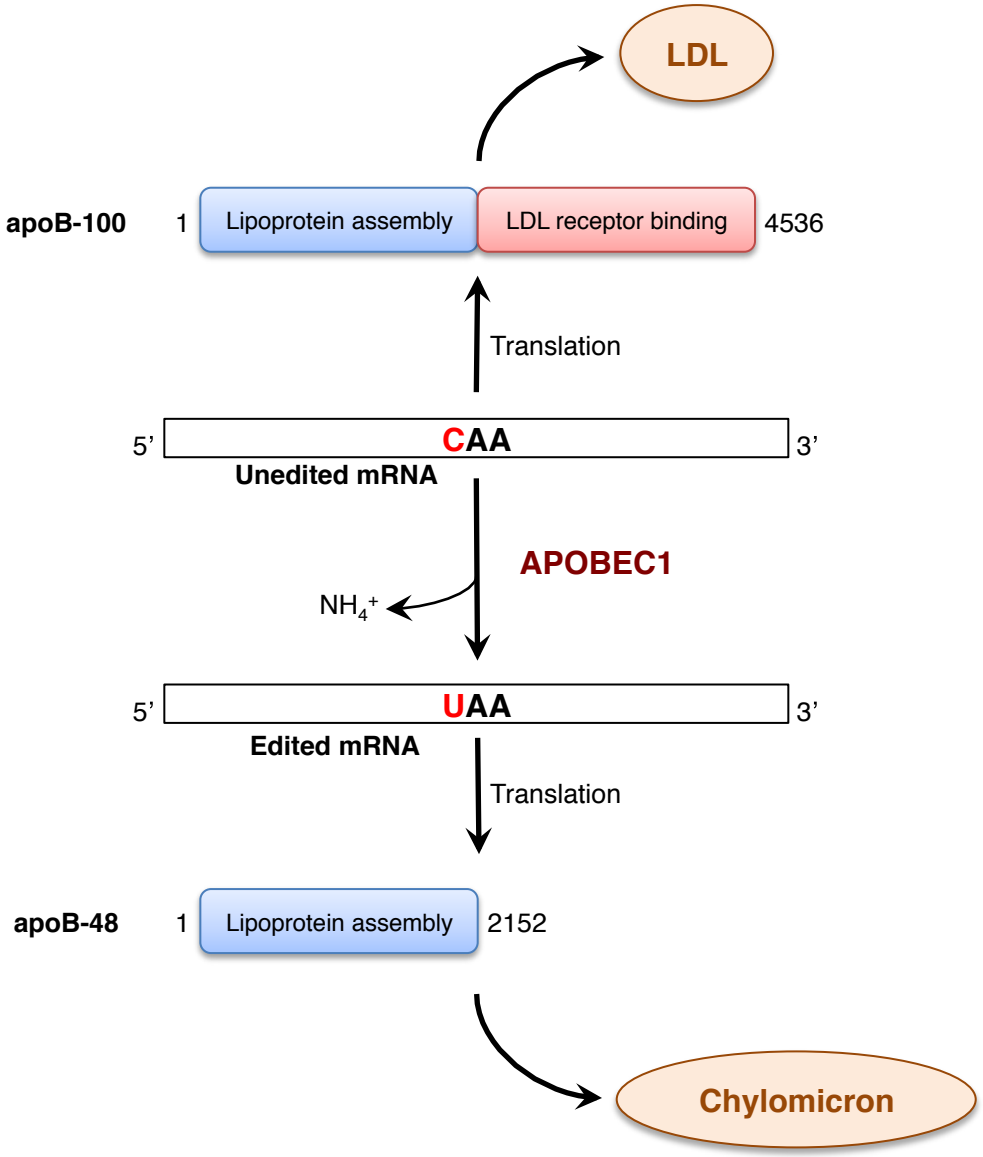


Figure 1.5. apoB mRNA editing by APOBEC1. apoB-100 and apoB-48 are derived from identical primary mRNA transcripts. In the absence of APOBEC1 editing, full-length apoB-100 is synthesized and incorporated into LDL lipoprotein particles. In small intestinal enterocytes, APOBEC1 site-specifically deaminates a cytidine in the apoB mRNA, thereby converting a glutamine codon (CAA) to a STOP codon (UAA). Upon translation, the truncated apoB-48 isoform is produced and incorporated into chylomicrons. Adapted from Stryer, L. *Biochemistry.* 5th edition, 2002.

The shorter apoB-48 isoform (designated as 48% of full-length apoB-100) is produced by the small intestine, in which it is essential for chylomicron formation and the absorption and transport of dietary lipid. ApoB is essential for mammalian life and development; homozygous apoB deletion mutants are embryonic lethal in mice (Farese et al., 1995). Both isoforms of apoB share identical primary transcripts and are not regulated by differential splicing or post-translational processing. In the small intestine, cytidine 6666 of the apoB mRNA is deaminated to uridine, thereby converting a glutamine codon (CAA) to a stop codon (UAA) and terminating elongation upon translation of apoB-48 (Chen et al., 1987; Powell et al., 1987). This site-specific mRNA modification is mediated by a multiprotein "editosome" complex, the catalytic component of which is APOBEC1 (Teng et al., 1993).

APOBEC1 is the only member of the AID/APOBEC family demonstrated to edit mRNA *in vivo*. The APOBEC1 protein contains a single zinc-dependent cytidine deaminase domain along with an RNA binding domain (MacGinnitie et al., 1995). In addition, the enzyme contains C-terminal protein-protein interaction motifs (Oka et al., 1997), which are thought to be involved in editing complex assembly. Evolutionarily, APOBEC1 is a relatively recent addition to the AID/APOBEC gene family; while AID and APOBEC2 are found from jawed fishes onward, APOBEC1 is present only in mammals (including marsupials) (Conticello et al., 2005). In humans, APOBEC1 expression appears to be exclusive to the small intestine (Lau et al., 1994). In many other mammals, however, while the small intestine remains the principal site of expression and apoB mRNA editing (Greeve et al., 1993), APOBEC1 can be found at additional tissue sites such as liver, spleen and lymph node (Hirano et al., 1997; Nakamuta

et al., 1995). *apobec1*$^{-/-}$ mice, though apparently healthy and fertile, do not edit apoB mRNA and have no apoB-48 in their serum (Hirano et al., 1996; Morrison et al., 1996).

The mechanism of apoB mRNA editing by APOBEC1 has been fairly well characterized. Though APOBEC1 shuttles between the cytoplasm and the nucleus (Bennett et al., 2006; Yang and Smith, 1997), apoB mRNA editing is an intranuclear event (Lau et al., 1991). Editing takes place either coincident with or immediately after transcript splicing and polyadenylation. APOBEC1 alone is not sufficient to recognize and deaminate apoB mRNA. Editing requires a large, multiprotein editosome complex, the minimal functional components of which are APOBEC1 and APOBEC1 Complementation Factor (ACF), an RNA binding protein. Purified ACF and APOBEC1 proteins are necessary and sufficient to support apoB editing *in vitro*.

ACF is a 64-kilodalton protein that was initially purified from baboon kidney extracts as an activity that functionally complements apoB mRNA editing by APOBEC1 (Mehta et al., 1996). The domain organization of ACF consists of three N-terminal RNA recognition motifs (RRMs), an arginine-glycine rich region, and a C-terminal dsRBM (Blanc et al., 2001a; Mehta et al., 2000). The tandem RRMs can form a large RNA binding surface capable of binding specific RNA sequences (Maris et al., 2005) . ACF also contains a nuclear localization signal that mediates its trafficking between the cytoplasm and nucleus (Blanc et al., 2003). This process is modulated by phosphorylation at multiple residues in ACF and contributes to the regulation of apoB mRNA editing in rodent hepatocytes (Lehmann et al., 2007; Lehmann et al., 2006). The ACF gene is alternatively spliced to generate at least 4 protein isoforms, which have different

capacities to support apoB editing (Dance et al., 2002; Sowden et al., 2004). Interestingly, deletions of ACF are embryonic lethal in mice, indicating that this RNA-binding protein serves additional functions beyond apoB mRNA editing (Blanc et al., 2005).

Though APOBEC1 and ACF constitute the minimal functional unit of apoB mRNA editing, the *in vivo* editosome is likely more complex. Multiple proteins that interact with APOBEC1, ACF, and/or RNA have been identified, several of which can modulate apoB editing. Glycine-arginine-tyrosine-rich RNA binding protein (GRY-RBP) is an hnRNP family member with significant homology to ACF, especially within its three similar RRM domains (Blanc et al., 2001b; Lau et al., 2001a). GRY-RBP inhibits apoB mRNA editing, likely through the binding and sequestration of ACF. CUG triplet repeat RNA binding protein 2 (CUGBP2) also contains RRMs, interacts with ACF and apoB mRNA, and inhibits APOBEC1 editing (Anant et al., 2001a). Other proteins that interact with APOBEC1 and enhance apoB mRNA editing have been described, such as APOBEC-1 Binding Protein-1 (ABBP-1) (Lau et al., 1997) and ABBP-2 (Lau et al., 2001b). The functional significance of these and additional yet-to-be-identified editosome component and associating proteins *in vivo* remain poorly understood. However, observations of APOBEC1 and/or ACF interacting factors supports a model in which apoB mRNA editing is carried out by a dynamic multimolecular complex with various regulatory components that modulate the reaction.

The APOBEC1 editing site in apoB mRNA is specified primarily by APOBEC1/ACF editosome recognition of local sequence elements in the apoB transcript. As an RNA binding protein, APOBEC1 exhibits a strong preference

for RNA sequences rich in A and U nucleotides (Anant et al., 1995; Navaratnam et al., 1995). Consistent with this observation, the transcript region containing the apoB editing site has a notably high AU content (70% A/U in a 101 nt window centered on the target cytidine). However, general AU content is not sufficient to convey precise targeting; specific sequence motifs are also required. The apoB mRNA contains an 11 nt "mooring sequence" (UGAUCAGUAUA) 5 nt downstream (3′) of the edited cytidine that serves as the principal determinant for apoB editosome targeting (Shah et al., 1991). Introducing mutations to the mooring sequence diminishes or eliminates editing, depending on which residues are altered. Furthermore, the experimental insertion of a mooring sequence downstream of a cytidine in heterologous AU-rich RNA is sufficient to induce editing by the APOBEC1/ACF complex, albeit at lower efficiency than observed in apoB RNA (Backus and Smith, 1991; Driscoll et al., 1993). Additional *cis*-acting elements in the apoB mRNA include a 5′ AU-rich efficiency sequence and a 3′sequence that modulates editing frequency. However, the minimal sequence of the apoB mRNA required for editing includes only the editing site, a short portion of the 5′ efficiency sequence, and the mooring sequence along with a spacer element to separate it appropriately from the target cytidine (Backus and Smith, 1992; Shah et al., 1991).

The sequence containing the editing site in apoB mRNA adopts a flexible imperfect hairpin secondary structure, which was phylogenetically predicted (Hersberger et al., 1999) and later confirmed by NMR spectroscopy (Maris et al., 2005). The target cytidine lies at the 5′ portion of the hairpin loop, with the mooring sequence starting in the 3′ portion of the loop and continuing into the stem duplex. In a current mechanistic model for apoB mRNA editing (Maris and

Allain, 2009), editosome-associated ACF recognizes and binds the apoB mooring sequence through its RRMs (Dance et al., 2002; Lellek et al., 2000; Mehta et al., 2000) and mediates the melting of the stem-loop structure (Maris et al., 2005), thereby generating a single-stranded RNA substrate amenable to APOBEC1 editing. The spacer element and editosome architecture are thought to ensure that the target cytidine is properly positioned in the cytidine deaminase active site. Following the editing reaction, the editosome remains associated with the apoB mRNA to facilitate its export from the nucleus and/or protect the edited transcript, which now contains a premature termination codon, from nonsense mediated decay (Chester et al., 2003).

Following the characterization of APOBEC1 as an RNA cytidine deaminase, numerous studies have investigated the possibility of additional functions and targets for the enzyme beyond its role in apoB editing. Attempts to identify additional mRNA editing targets have not yield many physiologically relevant substrates. Transgene-mediated hepatic overexpression of APOBEC1 causes editing of the eukaryotic translation initiation factor 4 (Eif4g2, also referred to as NAT1) mRNA (Yamanaka et al., 1997), but little evidence supports Eif4g2 as a target of APOBEC1 at endogenous levels. C-to-U editing in neurofibromatosis 1 (NF1) mRNA has been observed in peripheral nerve-sheath tumors and introduces a stop codon within the NF1 coding sequence (Mukhopadhyay et al., 2002; Skuse et al., 1996). The editing site is adjacent to a mooring sequence-like motif and likely represents a genuine APOBEC1 editing event. However, the editing observed was at relatively low levels and only occurred in a subset of patient samples. Thus, though NF1 may represent an

important APOBEC1 target in certain neurofibromatosis type I tumors, it likely does not represent a physiological target of editing in healthy, steady state tissue.

Though additional mRNA editing substrates have not been identified, APOBEC1 participates in several other biological processes independent of its deaminase activity. As an RNA-binding protein, APOBEC1 can bind to the AU-rich 3' UTRs of several transcripts to modulate their stability. This type of regulatory mechanism has been observed in APOBEC1 stabilization of the Cyclooxygenase-2 (COX-2) (Anant et al., 2004), c-myc (Anant and Davidson, 2000), and Cholesterol 7α-Hydroxylase (Cyp7a1) (Xie et al., 2009) mRNAs. Dysregulation of Cyp7a1 may be responsible for the increased susceptibility to gallstone disease of *apobec1*$^{-/-}$ mice. Such editing-independent functions may imply a broader biological influence for APOBEC1 than previously assumed.

1.4. Statement of the Problem

Multiple DNA and RNA editing enzymes, including APOBEC1 and AID, were identified based on prior observation of specific editing substrates. For AID, the recognition of point mutations in genes encoding antibody variable regions prompted a search for the DNA modification enzyme responsible. Similarly, the detection of an edited cytidine residue in apoB mRNA led to the discovery of APOBEC1. Though this approach has yielded many findings of great significance, it is limited in its potential to detect alternative biological roles for these enzymes beyond their initially observed editing substrates.

Several experimental observations suggest that APOBEC1 may have additional RNA editing targets aside from apoB. First, *apobec1*$^{-/-}$ phenotypes associated with binding and stabilization of transcript 3′ UTRs (Anant and Davidson, 2000; Anant et al., 2004; Xie et al., 2009) indicate that APOBEC1 interacts with many enterocyte mRNAs aside from apoB. Secondly, in many mammals, APOBEC1 is expressed in diverse tissues, including small intestine, liver, kidney, muscle and spleen (Nakamuta et al., 1995). Though transcribed and edited in intestine and liver, apoB mRNA is not present in tissues such as spleen. The absence of the apoB mRNA target raises questions as to the function of APOBEC1 in these tissues. Experimentation in our laboratory has detected APOBEC1 expression in specific immune cell types, including B cells, macrophages and dendritic cells. Furthermore, APOBEC1 expression in immune cells is differentially regulated in response to various stimuli (Figure 1.6). Finally, APOBEC1 exhibits significant positive selection throughout mammalian evolution (Sawyer et al., 2004), even though apoB mRNA editing is dispensable for viability. Though *apobec1*$^{-/-}$ mice are healthy and fertile (Hirano et al., 1996;

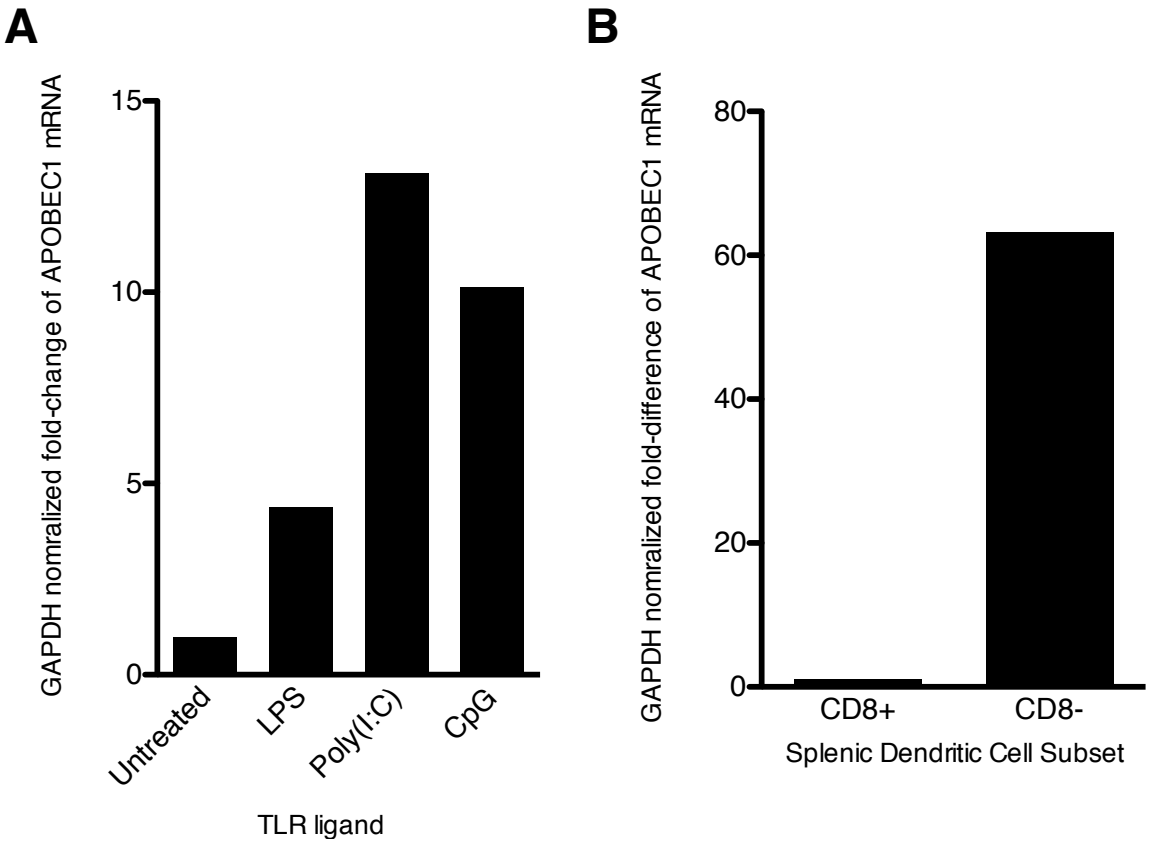


Figure 1.6. APOBEC1 mRNA is differentially regulated in immune cells. (A) Bone-marrow derived dendritic cells were stimulated with various TLR ligands (LPS, Poly(I:C), CpG oligonucleotides). GAPDH-normalized APOBEC1 mRNA levels were assessed by quantitative real-time PCR. (B) Splenocytes were harvested from C57/BL6 mice and FACSorted for $CD11c^{+}CD8^{+}$ and $CD11c^{+}CD8^{-}$ dendritic cell subsets. GAPDH-normalized APOBEC1 mRNA levels were assessed by quantitative real-time PCR.

Morrison et al., 1996), additional activities for APOBEC1 might be context-dependent. For example, APOBEC1 could function in the immune response to infection, as is the case for other AID/APOBEC family members.

Does APOBEC1 edit mRNAs other than the apoB transcript? As discussed above, previous attempts to identify additional editing sites by empirical and/or computational methods have not expanded the enzyme's target list (Smith et al., 2005). The likelihood of additional APOBEC1 RNA targets, as well as the discovery of other polynucleotide cytidine deaminases without known substrates (e.g. APOBEC2, APOBEC4), demands a novel approach to identify DNA and RNA editing events.

A *bona fide* RNA editing event can be defined as a discrepancy in the sequence of genomic DNA and RNA derived from the same cell or tissue sample. Similarly, an APOBEC1-dependent editing event would be present in samples derived from wild-type mice but absent from samples derived from congenic *apobec1*$^{-/-}$ mice. In essence, single nucleotide differences between two such RNA pools would represent APOBEC1 editing targets. However, without prior knowledge of which transcripts and sequences to examine, detecting such differences presents a challenge. Though biochemical techniques to detect nucleotide variations between pools of DNA (and RNA, by reverse transcription to cDNA) have been developed (Pan and Weissman, 2002), they are not sufficiently sensitive or practical for application at the scale of whole eukaryotic transcriptomes.

The recent development of ultra-high throughput sequencing technologies provides powerful tools for more comprehensive investigation of RNA editing. One recent study used target capture and deep sequencing to detect A-to-I

editing in numerous computationally predicted RNA targets (Li et al., 2009b). Whole transcriptome sequencing (RNA-Seq) represents another promising option for broad characterization of RNA editing. Though RNA-Seq is frequently used for transcriptome mapping and quantification (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Sultan et al., 2008), it has also been successfully applied to the analysis of single nucleotide polymorphisms (SNPs) in expressed genes (Chepelev et al., 2009; Heap et al., 2010). In considering a strategy for the discovery of APOBEC1 targets, I reasoned that given sufficient transcript coverage and read depth, the single nucleotide resolution of RNA-Seq could be used to identify candidate RNA editing sites throughout a transcriptome.

This thesis presents the development of an RNA-Seq method for the identification of mRNA editing events on a transcriptome-wide scale and its application to the discovery of APOBEC1 editing targets in small intestine enterocytes. This comparative screening approach involves ultra-high throughput sequencing of wild-type and *apobec1*$^{-/-}$ transcriptomes to identify APOBEC1-specific candidate editing sites. The small intestine provides an ideal experimental system for the development and validation of this technique, with apoB serving as an exceptional internal positive control. In addition to successful detection of the well-characterized apoB site, more than 30 previously undescribed editing events were identified and validated in small intestine enterocyte mRNA. Interestingly, these newly recognized editing sits are located in the 3′ UTRs of diverse transcripts. These sites share several characteristic sequence features, including a downstream (3′) motif similar to the mooring sequence in apoB mRNA. Many of the editing sites are located within transcript regions significantly conserved in mammalian evolution, implying possible

functional importance. These findings demonstrate the feasibility and utility of a transcriptomics approach to RNA editing studies and substantially expand the list of sites that undergo APOBEC1-dependent editing *in vivo*. In addition, the diverse editing targets identified suggest additional functions for APOBEC1 beyond its role in apoB-mediated lipid absorption and transport.

CHAPTER 2: PILOT STUDY – WHOLE INTESTINE

2.1. Experimental Strategy and Design

The empirical identification of RNA editing events can be achieved by detecting single nucleotide differences in a comparison of edited and unedited sequences. As such, assessing editing for a given mRNA transcript can be achieved by standard techniques such as reverse transcription, PCR and dideoxynucleotide Sanger sequencing. However, these techniques require prior knowledge of which transcript(s) to examine for editing. While the identification of unknown RNA editing events throughout a transcriptome raises numerous technical and analytical challenges, the basis for their recognition remains essentially the same: given a set of sequences exposed to editing activity and a corresponding set of sequences not exposed to editing activity, nucleotide variations are indicative of editing. Therefore, the experimental approach presented here begins with ultra-high throughput sequencing of whole transcriptomes isolated from wild-type and deaminase (editing enzyme)-deficient cells or tissues. Next, computational sequence analysis is used to identify single nucleotide variants specific to the wild-type transcriptome. Finally, with a defined list of candidates, RNA editing is validated by standard molecular biological techniques. This analysis workflow is presented in Figure 2.1. The details and rationale for this experimental strategy and its validation in mouse small intestine are described below.

2.1.1. Whole Transcriptome Sequencing

Recent technological advances now allow for the practical sequencing of entire eukaryotic transcriptomes. Ultra high-throughput short read sequencing

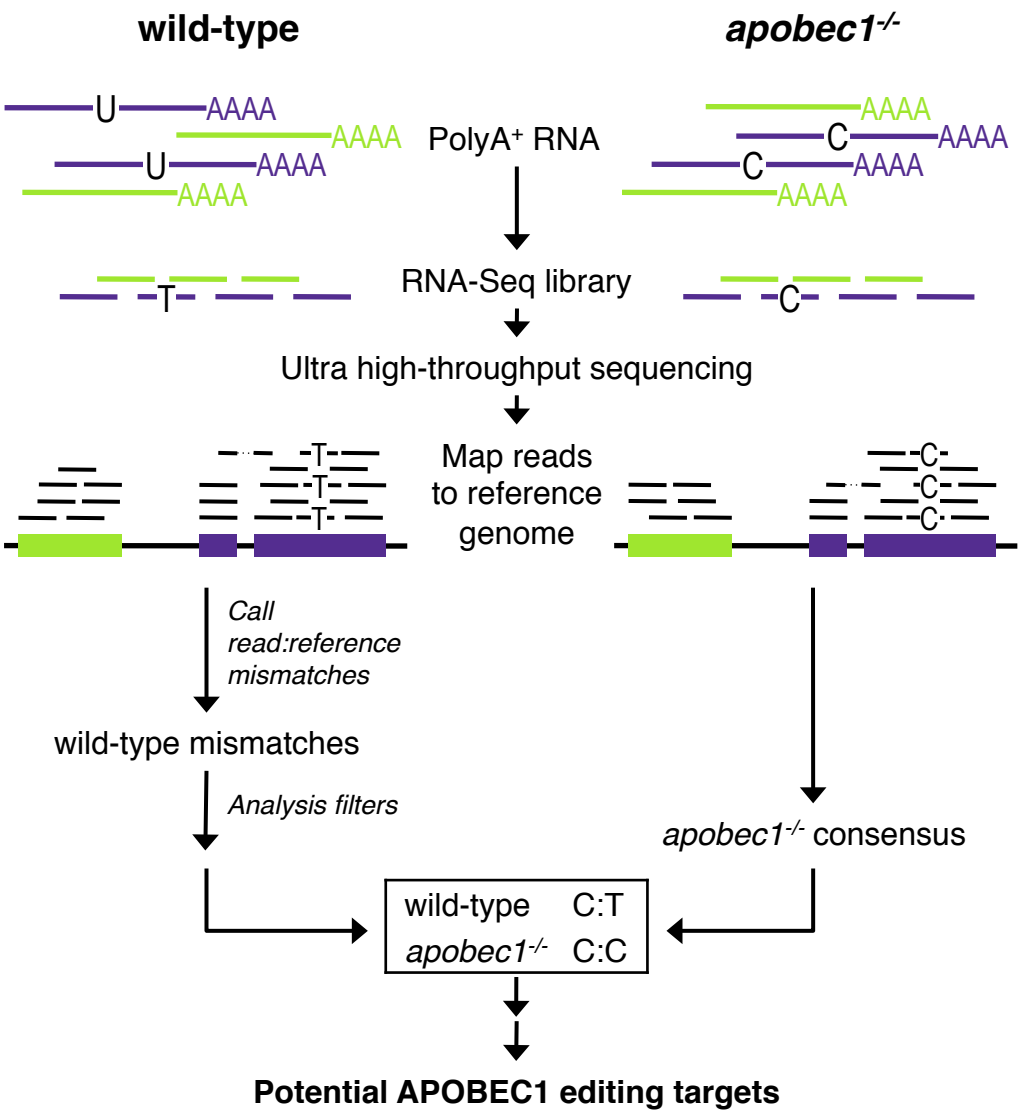


Figure 2.1. Comparative RNA-Seq screening strategy for the identification of APOBEC1 mRNA editing targets.

of cellular RNA via cDNA (RNA-Seq) is frequently used for gene expression profiling, but the nucleotide resolution of sequence data generated by this method can also be used to detect single base variants, such as exonic SNPs. As such, it should also be capable of detecting nucleotide variants generated by RNA editing.

As detailed below, RNA-Seq libraries are prepared by isolating polyA$^+$ mRNA, non-specifically fragmenting the mRNA into short (~200 nt) segments, and converting these to double-stranded cDNA duplexes containing flanking linker sequences. These linkers are recognized in the ultra-high throughput sequencing process, which is performed using the Illumina Genome Analyzer II. Sequencing generates tens of millions of short (36 nt) reads per library with an error rate of less than 1%. For each read, output files contain a unique identifier, sequence data and Phred-scaled quality scores for each base.

2.1.2. Mapping RNA-Seq reads to a reference genome

Following ultra-high throughput sequencing, reads are mapped and aligned to a reference genome. Mapping RNA-Seq reads presents a number of computational challenges. First, though sufficient for most transcripts of adequate quality (Mortazavi et al., 2008), reads 36 nt in length may not contain enough sequence information to map uniquely. This can present difficulties in mapping reads derived from orthologous and/or repetitive transcript regions. Secondly, sequencing libraries derived from randomly fragmented mature mRNAs will generate a significant proportion of reads spanning exon-exon junctions. Though such reads contain sequence contiguous in a mature, spliced mRNA transcript, the separated genomic context of distinct exons effectively

divides such sequences to at least two locations in the reference. Therefore, any RNA-Seq mapping algorithm must incorporate a strategy for the alignment of exon junction-spanning reads, which can constitute approximately 3-15% of mappable reads in a typical short (25 nt - 36 nt) read dataset (Mortazavi et al., 2008; data not shown). Finally, and of particular significance for RNA editing studies, many reads will contain incorrect base calls as a result of the inherent error of ultra-high throughput sequencing. A mapping strategy should be permissive for some read-to-reference mismatches resulting from sequencing error or RNA editing. Editing mismatches can be distinguished from sequencing errors in downstream analysis.

There are several academic and commercial software packages available for alignment of RNA-Seq short read data. Though many are likely compatible with a similar workflow, the "Tuxedo Tools," TopHat (for mapping RNA-Seq reads) (Trapnell et al., 2009) and Bowtie (for aligning reads to reference genome) (Langmead et al., 2009) were selected on account of their flexibility and efficiency. Taking into account the challenges described above and the demands of detecting nucleotide variations indicative of mRNA editing, the following alignment strategy was implemented:

a. **Alignments should allow for mismatches in reads relative to the reference genome.**
b. **Alignments should be "quality conscious."** Due to the relatively high error rate of ultra high-throughput sequencing and relatively low probability of an mRNA editing event, base quality scores should be taken into account for mapping and mismatch calling algorithms.

c. **Alignments should be unique.** As an mRNA editing event is detected as a read mismatch to reference, it is imperative that mismatches occur only at "real" editing sites and not as a result of sequencing errors. Therefore, the mapping algorithm should suppress all reads for which an alignment is not "unique", i.e. it can be satisfactorily mapped to more than one location in the genome while still satisfying mismatch and quality limits. Though such an approach will dramatically reduce the number of potentially "good" alignments, such stringency ensures high confidence in mismatch hits.

d. **Reads spanning exon-exon junctions should be mapped accordingly.** Different alignment algorithms approach the problem of mapping reads derived from mature, spliced mRNAs to genomic reference with various strategies; these include alignment to an artificial "splice-ome" reference sequence of all predicted exon-exon junctions (Mortazavi et al., 2008), and *ab initio* mapping of reads to junctions predicted by read distribution and pileup (Trapnell et al., 2009). In order to allow for read mapping across novel splice junctions not present in annotation databases, the TopHat algorithm attempts to identify exon borders based on reads derived entirely from individual exons. Briefly, reads are first mapped to the reference genome irrespective of splicing information. Those reads (including junction-spanning reads) that cannot be mapped are set aside. The alignment pattern of mappable reads is then used to identify exons and place them within probable gene models, which are then used to generate different combinations of exon-exon junctional reference sequence. Next, the previously

unmappable reads are aligned against the newly-generated junctional reference. Junctional read alignments are then split accordingly and output with genomic coordinates. This *ab initio* mapping algorithm can also be supplemented with previously characterized gene annotations, allowing for alignments to known and novel transcript models.

Finally, all TopHat alignments are output to a single comprehensive Sequence Alignment/Map (SAM) file, which is used in downstream mismatch analysis.

2.1.3. Identifying read-to-reference mismatches

Once suitable RNA-Seq alignments are generated, single nucleotide mismatches in wild-type reads relative to the reference sequence are identified. There are few (if any), software options specifically designed to call mismatches generated by RNA editing events. However, several analysis packages incorporate genomic SNP calling algorithms, which can also be implemented for RNA mismatch analysis. Most importantly, as for the alignment step, mismatch calling should be "quality conscious" to ensure high confidence in read:reference discrepancies.

The publicly available, open source SAMTools software package developed at the Sanger Institute was selected for mismatch analysis (Li et al., 2009a). Beginning with a TopHat generated SAM file, SAMTools generates a "pileup" file of aligned RNA-Seq reads relative to the reference genome. Unlike the input SAM file, in which entries are organized by individual reads, the pileup file contains reads and their qualities on a reference base-by-base scale

and a corresponding consensus base call at each position (Figure 2.2). Thus, for a given nucleotide position in the transcriptome, the pileup file includes data on the number of reads covering the position, the base content and corresponding quality score of each read, the consensus base derived from the aggregate reads at the position, and a consensus probability score indicating the likelihood of an incorrect consensus base call. Furthermore, when the base content of reads mapped to given a position do not match the reference base, the pileup file contains a consensus mismatch base call and corresponding mismatch probability score. The consensus probability score is defined as the Phred-scaled probability that the consensus base call is incorrect, while the mismatch probability score is defined as the Phred-scaled probability that the consensus base call matches the reference base (Li et al., 2008). Both values are based on read depth and read quality scores.

In order to generate a starting list of candidate RNA editing sites, consensus and mismatch information is used to identify significant single nucleotide read:reference mismatches from the wild-type dataset. However, the list of variations may contain large numbers of read:reference mismatches unrelated to mRNA editing. These mismatches may be a result of genomic SNPs, sequencing errors, misaligned reads, reverse transcription/amplification errors and unrelated mRNA modification processes. As such, the initial variation list must be filtered on several criteria appropriate to RNA editing, including error probability, sequence type (exons of known or predicted genes only), read:reference mismatch base calls, known SNPs and repetitive elements. A particular series of filters will be unique to the mRNA editing enzyme studied; for cytidine deaminases, candidate sites should be limited to T:C read:reference

Chromosome	Position	Reference Base	Read Consensus	Consensus Score	Mismatch Score	Read Depth	Read Bases	Read Quality Scores
chr1	192830760	T	T	66	0	13	,,.,,,,,,,,,,	IIIIIIIIIIII
chr1	192830761	G	R	35	95	14	aaAaaaa,a,a,aa	IIIIIIIIIIIII
chr1	192830762	A	A	69	0	14	,,.,,,,,,,,,,,	IIIIIIIIIIIII
chr1	192830763	A	A	69	0	14	,,.,,,,,,,,,,,	IIIIIIIIIIIII
chr1	192830764	A	A	69	0	14	,,.,,,,,,,,,,,	IIIIIIIIIIIII
chr1	192830765	C	C	66	0	13	,.,,,,,,,,,,,	IIIIIIIIIIII
chr1	192830766	A	A	63	0	12	.,,,,,,,,,,,	IIIIIIIIIII
chr1	192830767	G	G	63	0	12	.,,,,,,,,,,,	IIIIIIIIIII
chr1	192830768	A	A	63	0	12	.,,,,,,,,,,,	IIIIIIIIIII
chr1	192830769	A	A	63	0	12	.,,,,,,,,,,,	IIIIIIIIIII

Figure 2.2. Excerpt from a pileup file. RNA-Seq read data are represented at individual base positions in the reference genome. This excerpt depicts read data for chr1:192830760 – 192830760. For read bases, periods or commas indicate matches to the reference sequence. Read quality scores are represented in ASCII character format; numerical scores are equal to (ASCII code) – 33. A consensus G-to-A mismatch is apparent at chr1:192830761.

mismatches. Though stringent filters could potentially discard genuine RNA editing sites, they also minimize the screen's false detection rate and were deemed necessary for a manageable analysis.

2.1.4. Identifying specific mRNA editing sites

Once read:reference mismatches in the wild-type dataset are identified and filtered, specific editing events are distinguished from other variants. Mapped and aligned RNA-Seq reads from the deaminase-deficient sample serve as a control for this purpose. Each wild-type mismatch site is compared to the corresponding site in the deaminase-deficient pileup dataset; if the deaminase-deficient alignment also contains the read:reference mismatch, the variant is not editing-specific and discarded. However, if the deaminase-deficient reads (and corresponding consensus) do not contain a mismatch, the site is selected as a candidate mRNA editing site. The reliability of this comparison depends on sufficient read coverage and base quality in both the wild-type and deaminase-deficient datasets.

2.1.5. Validation of candidate mRNA editing sites

With a defined list of candidate sites provided by the RNA-Seq screen, mRNA editing can be confirmed using standard techniques. First, genomic DNA from each library is sequenced to rule out heterozygosity. Next, genomic DNA sequences from wild-type and deaminase-deficient samples are compared to rule out a SNP. Wild-type cDNA sequence is compared to wild-type genomic DNA sequence to confirm mRNA editing. Finally, deaminase-deficient cDNA

sequence is compared to wild-type cDNA sequence to validate specific deaminase activity.

2.1.6. Testing the screening method: APOBEC1 in the small intestine

Developing a novel experimental approach to the identification of RNA editing targets requires technical validation by means of a positive control. As such, the feasibility and utility of the RNA-Seq screen was tested on APOBEC1 editing in the small intestine. ApoB mRNA is the only confirmed target of APOBEC1 in the intestine. In wild-type mice, editing of the apoB mRNA is efficient; approximately 90-100% of intestinal apoB transcripts are site-specifically edited from cytidine to uridine (Figure 2.3). In contrast, no editing is detectable in apoB transcripts isolated from *apobec1*$^{-/-}$ intestine. Therefore, apoB represents an unambiguous positive control for mRNA editing. If the RNA-Seq screening approach is to be useful in identifying unknown, potentially inefficiently edited targets, it should be able to detect apoB editing using the same workflow.

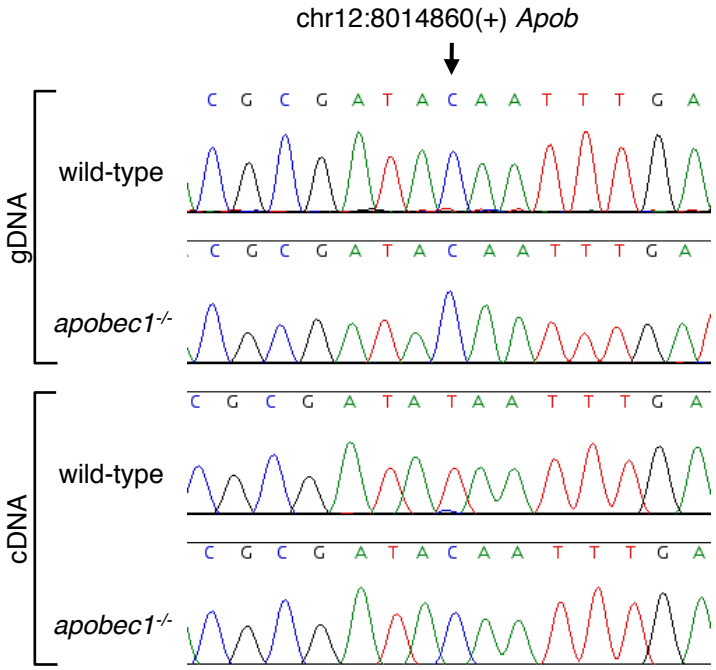


Figure 2.3. ApoB mRNA editing in the small intestine. Sanger sequencing of genomic DNA and cDNA derived from jejunal segments of wild-type and *apobec1*$^{-/-}$ mice illustrates APOBEC1-dependent apoB mRNA editing. Editing is site-specific and efficient; near complete C-to-U(T) conversion is apparent in the wild-type cDNA chromatogram.

2.2. Experimental Procedures

2.2.1. Mice

C57/BL6 wild-type and congenic *apobec1*$^{-/-}$ mice were used at 6-8 weeks of age. *Apobec1*$^{-/-}$ mice (Hirano et al., 1996) were generously provided by N. Davidson (Washington University School of Medicine, St Louis, MO).

2.2.2. Preparation of RNA-Seq Libraries

The RNA-Seq library preparation protocol was adapted from (Mortazavi et al., 2008) and Illumina product literature.

Total RNA was isolated from whole intestine tissue by organic extraction with TRI Reagent (Ambion). The starting total RNA was confirmed to be of high quality by Agilent Bioanalyzer 2100 analysis. PolyA$^+$ mRNA was isolated using the MicroPoly(A)Purist Kit (Ambion). Total RNA was incubated on polyT resin at room temperature for 1 hr to ensure maximum binding and recovery of polyA$^+$ mRNA. Following initial polyA$^+$ mRNA enrichment, the procedure was repeated with fresh polyT resin to maximize depletion of non-mRNA species. Purity and size distribution of enriched polyA$^+$ mRNA was monitored by Agilent Bioanalyzer 2100 analysis.

Following confirmation of quality, the polyA$^+$ mRNA was concentrated by ethanol precipitation and resuspended in water to a final concentration of 100 ng/ul. Approximately 800 ng of polyA$^+$ mRNA was then non-specifically fragmented by supplementing with fragmentation buffer (final composition: 40 mM Tris acetate, pH 8.2, 100 mM potassium acetate, 30 mM magnesium acetate) and incubating at 94^{o}C for 4 min 30 sec. Fragmented polyA$^+$ mRNA was then

washed and concentrated by ethanol precipitation. The distribution of mRNA fragment sizes was evaluated by the Agilent Bioanalyzer 2100.

First-strand cDNA was prepared using Superscript III Reverse Transcriptase (Invitrogen) with random hexamer priming. The reverse transcription reaction was incubated at 51°C for 45 min prior to enzyme inactivation at 70°C. Second-strand synthesis was performed using the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen) and its RNase H, *E. coli* DNA polymerase and *E. coli* DNA ligase components. Following the 2 hr synthesis reaction, T4 DNA polymerase was added to fill-in offset ends.

In order to facilitate subsequent linker oligonucleotide ligation, adenine overhangs were added to the ds-cDNA by Klenow Fragment (3´→5´ *exo*$^{-}$) in the presence of 200 mM dATP (incubated at 37°C for 30 min). Next, Illumina sequencing adaptors were ligated to the ds-cDNA duplexes using concentrated T4 DNA ligase (New England Biolabs).

RNA-Seq ultra high-throughput sequencing is most effective with cDNA fragments of relatively short (~200 nt) and defined size. As RNA fragmentation can be variable and somewhat heterogeneous, libraries were separated by agarose gel electrophoresis and extracted based on DNA size standards. Though sample lanes were not visible due to the small amount of cDNA, alternating DNA size marker lanes served as a guide by which to excise gel slices containing cDNA duplexes of 200 +/- 25 nt. Following spin-column gel extraction (Qiagen), RNA-Seq libraries were prepared by amplifying (15 cycles PCR) the adaptor-ligated ds-cDNA with Phusion high-fidelity DNA polymerase (New England Biolabs) and Illumina primers (PE 1.0 and PE 2.0) complimentary to portions of

the ligated adaptor sequences. The concentration and purity of final amplified RNA-Seq libraries was determined by Nanodrop spectrophotometer and Agilent Bioanalyzer 2100.

2.2.3. Ultra high-throughput sequencing

Ultra high-throughput sequencing on the Illumina Genome Analyzer II (GAII) was performed with Scott Dewell (Rockefeller University Genomics Resource Center) and is detailed in the corresponding Illumina technical literature (Illumina, 2008a, b). A brief summary of the technique as it applies to this project appears here.

RNA-Seq libraries were diluted to concentrations ranging from 4 to 8 pM and hybridized to an Illumina GAII flowcell, on which covalently linked oligomers (complimentary to the library adapter sequences) capture the ds-cDNA templates. Libraries were then bridge-amplified onto flowcell-bound oligomers and removed by denaturing and washing. Bridged templates were polymerase amplified, thereby generating flowcell-bound, sequence-matched "clusters." Following cluster "clean up" to standardize strand polarity, ultra-high throughput sequencing was performed by Illumina sequencing-by-synthesis reaction. The sequencing reaction consists of stepwise cycles that proceed iteratively; total cycle number dictates read length. For the samples described here, 36 cycles were used to generate reads 36 nt in length. Polymerase extends the 3' end of primer (annealed in the final amplification step) through incorporation of fluorescently labeled (by base), reversibly terminating nucleotides. After the incorporation of one chain-terminating nucleotide, the entire flowcell is imaged in small "tiles," 120 tiles per lane. Four images are

captured per tile, one for each nucleotide laser/filter combination. A cleavage step removes the fluorophore and reverses chain termination. After a wash step, the cycle is repeated until the desired read length has been reached.

Raw image data were processed using the standard Illumina software pipeline (SCS2.4). Real-time analysis and basecalling generated files containing data on each sequencing read (*qseq.txt), intensities (*.cif) and noise profiles (*.cnf). Due to the nature of the sequencing methodology, certain corrective measures are implemented to account for spectral crosstalk from the fluorophores and the fact that not all strands in each cluster are extended in perfect synchrony; some will not be extended and others will be extended by two bases at certain cycles, leading to what is termed "phasing." Spectral crosstalk and phasing were corrected using the Bustard (GA Pipeline 1.4.0) program and the phiX174 control lane as recommended by Illumina. The resulting *qseq.txt files contain read IDs, sequence, and quality scores for each flowcell tile. Standard FASTQ files were generated using the Gerald (GA Pipeline 1.4.0) program.

2.2.4. Mapping RNA-Seq reads to the reference genome

Prior to mapping, RNA-Seq reads were examined for overall quality and base content. Present analysis and prior evidence indicated that priming reverse transcription with random hexamer primers leads to an overrepresentation of G and C residues in the initial sequencing cycles. To eliminate this potential source of error, prior to mapping, each short read was "trimmed" – the 5′ first two bases (and associated quality information) were digitally removed from the FASTQ

data files with the FASTQ/A Trimmer tool, part of the FASTX Toolkit software package (available at *http://hannonlab.cshl.edu/fastx_toolkit/index.html*).

The rationale and strategy for RNA-Seq read mapping is detailed above. Based on the mapping criteria described, reads were mapped to the C57/BL6 reference mouse genome (NCBI37/mm9) using the Tuxedo Tools software packages, Bowtie (short read alignment) (Langmead et al., 2009) and TopHat (spliced read mapping for RNA-Seq) (Trapnell et al., 2009). A GFF file of RefSeq gene models was supplied to supplement exon-exon junction mapping. The command line parameters for mapping and alignment were as follows:

$ tophat –n 2 -g 1 --segment-length 34 -a 12 -m 1 -p 8 –G <RefSeq.gff3> <mm9_genome> <Read_Files.fastq>

-n 2 allow for up to 2 mismatches to reference in seed region (first 28nt); quality conscious

-g 1 suppress all alignments for reads that map to >1 location in reference

--segment-length 34 read length is 34 nt for alignment

-a 12 for exon-exon junction reads, require at least 12 bases (“anchor”) on either side of junction

-m 1 for exon-exon junction reads, allow for up to 1 mismatch in anchor segment

-p 8 use 8 processor cores for computations

-G supplement splice junction mapping with supplied gene model annotation file <RefSeq.gff3>

The resulting SAM output files were used in downstream mismatch analysis and "wiggle" plot output files were visualized with the UCSC genome browser to evaluate read coverage of expressed transcripts.

RNA-Seq read analysis, trimming, mapping and alignment were performed on virtual server instances provided by the Amazon Elastic Compute Cloud (EC2) (*http://aws.amazon.com/ec2/*) and running Ubuntu Linux version 8.0 or above.

2.2.5. Transcriptome sequence analysis

According to the analysis strategy detailed above, single nucleotide read:reference mismatches were identified with the SAMTools software package (Li et al., 2009a). The standard SAMTools workflow was used to generate a pileup output that contained all mapped RNA-Seq reads and their quality scores on a reference base-by-base scale and a corresponding consensus base call at each position. This information was used to identify single nucleotide variants in the wild-type dataset with a quality-conscious algorithm (Li et al., 2008).

The index of read:reference variant sites was filtered on several criteria to restrict analysis to those mismatches related to APOBEC1 editing. First, as many mismatch sites are the result of off-target mapping to intergenic and intronic sites, only those sites that mapped to RefSeq exons (UCSC Table Browser, NCBI 37/mm9) were retained. Next, to identify sites consistent with APOBEC1 cytidine deaminase activity, only sites at which the reference base was a C and the read consensus call included T were selected for additional consideration. Mismatch sites annotated as SNPs (dbSNP build 128, available at ftp.ncbi.nih.gov/snp) were also discarded. Finally, the remaining sites were

compared to read consensus base calls in the *apobec1*$^{-/-}$ dataset. Only those sites at which wild-type read consensus contained C:T mismatches and *apobec1*$^{-/-}$ read consensus contained a high-confidence C:C match were deemed potential editing sites. This list was further reduced by removing those sites with insufficient read depth (<5 reads for wild-type, <3 reads for *apobec1*$^{-/-}$) and/or insufficient confidence scores (Phred-scaled mismatch probability < 45 for wild-type, Phred-scaled consensus probability < 30 for *apobec1*$^{-/-}$).

Unless otherwise described, all analysis filters and database queries were performed with standard shell scripts or the Galaxy genomics web portal (*http://g2.bx.psu.edu*) (Taylor et al., 2007).

2.2.6. APOBEC1 editing site validation

Genomic DNA and RNA were prepared from wild-type and *apobec1*$^{-/-}$ small intestine by standard methods. cDNA was prepared from total RNA by Superscript III reverse transcriptase (Inivitrogen) and random hexamer priming. Sequences containing potential APOBEC1 editing sites were PCR amplified using TurboPfu high-fidelity polymerase (Stratagene). Primer extension sequencing was performed by GENEWIZ, Inc. (South Plainfield, NJ) using Applied Biosystems BigDye version 3.1. The reactions were then run on Applied Biosystem's 3730xl DNA Analyzer.

2.3. Results

2.3.1. RNA-Seq screen for APOBEC1 mRNA editing sites - small intestine

To test the efficacy and feasibility of the proposed experimental approach for identifying deaminase-specific mRNA editing sites, the comparative RNA-Seq screen was applied to the small intestine transcriptome. Jejunal segments were isolated from C57/BL6 wild-type and congenic *apobec1*$^{-/-}$ mice. RNA-Seq libraries were prepared from polyA^{+} mRNA and sequenced by ultra-high throughput methods. As a pilot study, a relatively low number of reads were acquired: 13,698,876 reads for wild-type, 15,040,776 reads for *apobec1*$^{-/-}$. Reads were mapped to the mouse reference genome (mm9, NCBI 37.1), allowing for up to two mismatches per sequence. Only reads with sufficient quality scores that mapped to unique sites in the genome were used for analysis (8,297,959 reads for wild-type, 8,365,157 reads for *apobec1*$^{-/-}$). A qualitative analysis of read distribution indicated that, while not comprehensive, read coverage of transcripts at moderate to high expression levels was extensive and somewhat evenly distributed (Figure 2.4). Gaps in otherwise well-covered transcripts were usually a result of repetitive or low-complexity local sequence content.

The strategy for detecting potential RNA editing events described above was applied to the mapped read datasets. Sites within RefSeq exons at which the reference genome contained a C and wild-type reads contained Ts (or reference G and RNA-Seq read As for (-)-strand transcripts, in genomic context) were compared to corresponding *apobec1*$^{-/-}$ reads. Sites at which the *apobec1*$^{-/-}$ read consensus matched the reference sequence were considered for additional analysis. After filtering out those sites with insufficient read coverage and/or

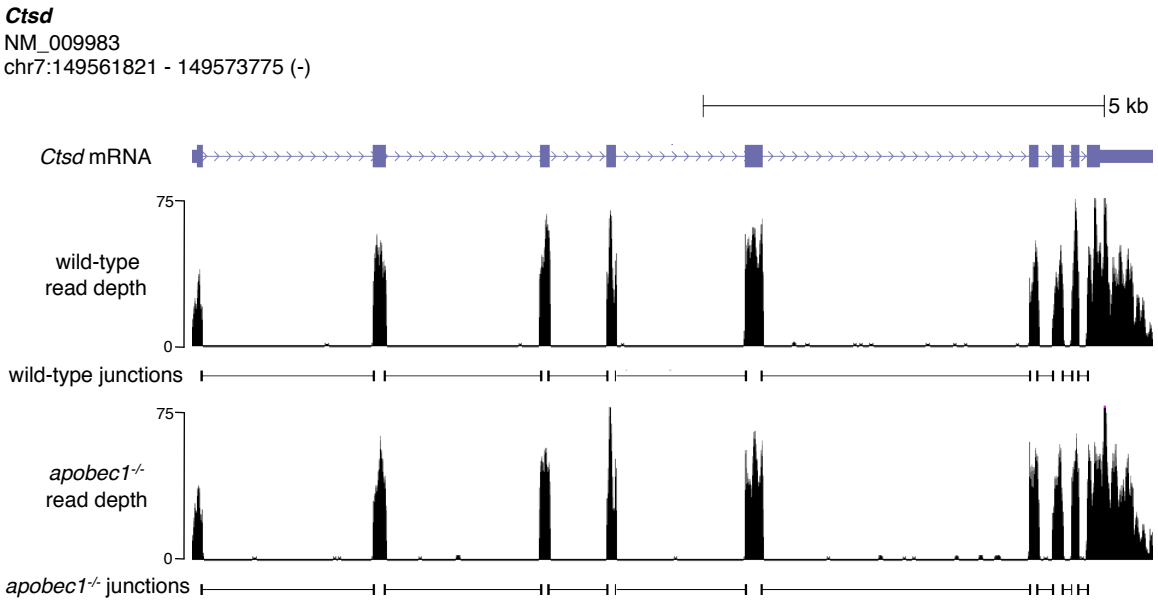


Figure 2.4. RNA-Seq read coverage of a transcript expressed at moderate levels in small intestine. The *Ctsd* gene encoding cathepsin D serves as a representative example for transcript read coverage. As RNA-Seq libraries are prepared from polyA$^+$ mRNA, reads map predominantly to exons (thick blue bars). Read coverage and mapped exon-exon junctions are similar in wild-type and *apobec1*$^{-/-}$ datasets.

mismatch/consensus probability scores, 35 remaining sites were designated candidate APOBEC1 editing targets (Table 2.1). When sites were ranked by mismatch probability score and read depth, the top hit was the well-characterized editing site in apoB mRNA. This result confirmed that the sequencing methodology and analysis pipeline can successfully detect single nucleotide editing events on a transcriptome scale.

Though the detection of editing in apoB mRNA was anticipated, the significance of the other 34 candidate editing sites indentified in the screen was somewhat unclear. While the various analysis filters were designed to minimize mismatches unrelated to specific RNA editing activity, some false positive hits were expected. Sample-specific false positives could be caused by sequencing errors, PCR amplification of polymerase errors, unannotated SNPs or read mapping artifacts. Alternatively, the candidate sites could represent genuine, previously undescribed APOBEC1 RNA editing targets.

These possibilities were explored by additional examination of RNA-Seq data and validation experiments. For each candidate site, individual read alignments were assessed for base content and potential artifacts. As expected based on filter design, mismatches were consistent with cytidine deaminase activity; if distinct from reference, variant bases were almost always T (or A for (-)-strand transcripts). There was no evidence of non-specifically mixed nucleotide mismatches. Furthermore, reads covering candidate sites were often staggered, with a range of distinct start and end coordinates. This diminishes the likelihood that mismatches were caused by PCR amplification of revere transcriptase or polymerase errors during library preparation, as they would be expected to appear as identical reads. However, unlike the apoB target site, at which greater

Table 2.1. Candidate APOBEC1 editing sites (small intestine)

			wild-type					*apobec1*$^{-/-}$				
Genome Site	Gene	Ref. Base	Read Cons.	P Cons.	P Mism.	Read Depth	Editing Freq.	Read Cons.	P Cons.	P Mism.	Read Depth	Editing Freq.
chr12:8014860(+)	Apob	C	T	255	255	266	0.94	C	255	0	200	0.00
chrX:109671648(+)	2010106E10Rik	C	Y	228	228	181	0.43	C	254	0	99	0.00
chr8:46391931(-)	Cyp4v3	G	R	228	228	105	0.32	G	129	0	34	0.00
chr2:121978638(+)	B2m	C	Y	228	228	323	0.26	C	255	0	173	0.00
chr16:84955113(-)	App	G	R	228	228	82	0.28	G	250	0	74	0.00
chrX:136207009(+)	Rnf128	C	Y	216	216	167	0.22	C	219	0	87	0.00
chr3:129616676(+)	Casp6	C	Y	193	193	56	0.25	C	167	0	59	0.00
chr3:73442586(-)	Bche	G	R	191	191	33	0.33	G	57	0	10	0.00
chr18:24445094(+)	Galnt1	C	Y	107	107	12	0.42	C	60	0	11	0.00
chr15:99239051(+)	Tmbim6	C	Y	107	107	90	0.18	C	255	0	104	0.00
chr3:73442584(-)	Bche	G	R	106	106	34	0.24	G	36	0	13	0.00
chr12:38308269(+)	Tmem195	C	Y	104	104	13	0.38	C	48	0	7	0.00
chr5:87984364(-)	Sult1d1	G	R	102	102	15	0.33	G	48	0	7	0.00
chr10:57235791(-)	Serinc1	G	R	96	96	19	0.32	G	57	0	10	0.00
chr1:192830761(-)	Mfsd7b	G	R	35	95	14	0.79	G	69	0	14	0.00
chr17:44416335(+)	Clic5	C	Y	86	86	29	0.21	C	93	0	22	0.00
chr12:88650627(-)	Sptlc2	G	R	83	83	13	0.31	G	60	0	11	0.00
chrX:106355759(+)	Sh3bgrl	C	Y	79	79	13	0.31	C	48	0	7	0.00
chr13:96397211(-)	Iqgap2	G	R	74	74	9	0.44	G	51	0	8	0.00
chr13:96397404(-)	Iqgap2	G	R	67	67	12	0.33	G	57	0	10	0.00
chr17:43611473(-)	Mep1a	G	R	62	62	17	0.24	G	42	0	5	0.00
chr18:6789843(+)	Rab18	C	Y	61	61	9	0.33	C	42	0	5	0.00
chr3:73443602(-)	Bche	G	R	60	60	14	0.29	G	42	0	5	0.00
chr1:90115552(+)	Ugt1a5	C	Y	60	60	22	0.23	C	75	0	16	0.00
chr14:79981748(+)	Elf1	C	Y	56	56	8	0.38	C	69	0	14	0.00
chr14:73762178(-)	Itm2b	G	R	56	56	7	0.43	G	51	0	8	0.00
chrX:106356686(+)	Sh3bgrl	C	Y	55	55	33	0.21	C	105	0	26	0.00
chrX:109671857(+)	2010106E10Rik	C	Y	54	54	29	0.21	C	66	0	13	0.00
chr7:96290044(-)	Tmem135	G	R	53	53	8	0.38	G	42	0	5	0.00
chr2:109739674(+)	Lin7c	C	Y	53	53	16	0.25	C	99	0	24	0.00
chr1:4835552(+)	Lypla1	C	Y	53	53	31	0.16	C	65	0	26	0.00
chr5:65805030(-)	Ugdh	G	R	48	48	45	0.13	G	123	0	44	0.00
chr12:38308281(+)	Tmem195	C	Y	48	48	45	0.13	C	87	0	20	0.00
chr11:20125336(+)	Rab1	C	Y	46	46	34	0.18	C	120	0	31	0.00
chr11:109313859(-)	Slc16a6	G	R	46	46	14	0.21	G	54	0	9	0.00

than 90% of the wild-type RNA-Seq reads contained a T instead of the reference C, wild-type samples at the other candidate sites contained smaller proportions of mismatch-containing reads.

2.3.2. Validation of candidate editing sites in small intestine

In order to determine if the candidates identified in the RNA-Seq screen could represent actual targets of APOBEC1 editing, several sites were selected for experimental validation. Standard dideoxynucleotide Sanger sequencing was used to examine the sites in genomic DNA and cDNA derived from murine intestine. All validation samples were independently prepared from different mice than those used for RNA-Seq libraries. Evidence of APOBEC1-dependent C-to-U(T) mRNA editing was apparent in 15 of 16 sites examined. Representative examples of sequencing chromatograms appear in Figure 2.5. The C/T chromatogram peaks from wild-type cDNA were of varied intensities, with C as the dominant peak in most sequences. This observation is consistent with the low fraction of mismatch-containing wild-type RNA-Seq reads. Nonetheless, these experiments confirm that APOBEC1 edits additional mRNAs other than the apoB transcript and that these sites can be identified by the comparative RNA-Seq screening method described here.

The recognition of additional targets for APOBEC1, an enzyme long thought to be monospecific for apoB mRNA, raises many additional questions regarding the enzyme's specificity, efficiency and physiological functions. Though this pilot study demonstrated proof of principle for both the RNA-Seq screening approach and APOBEC1-specific editing beyond apoB, it had several limitations. First, though sufficient for good coverage of transcripts expressed at

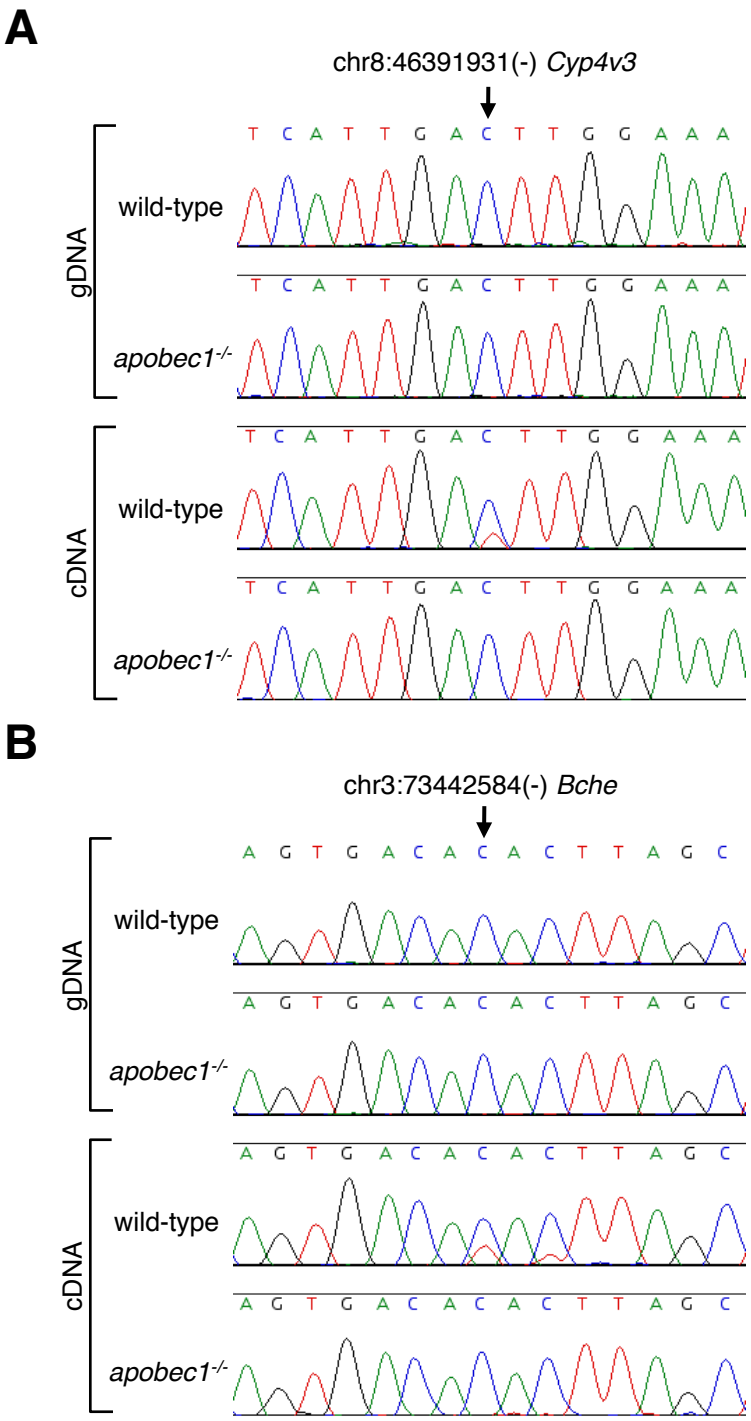


Figure 2.5. Validation of candidate APOBEC1 editing sites in whole intestine. Representative examples of conventional Sanger sequencing for whole intestine genomic DNA (gDNA) and mRNA (reverse-transcribed to cDNA) are shown. C-to-U(T) editing is apparent only in wild-type cDNA. (A) chr8:46391931(-) in the *Cyp4v3* transcript, and (B) chr3:73442584(-) in the *Bche* transcript.

relatively high levels, the number of RNA-Seq reads generated by ultra-high throughput sequencing was inadequate for extensive single-nucleotide coverage of the transcriptome. Next, a considerable proportion of RNA-Seq reads were likely generated from transcriptomes irrelevant for APOBEC1 editing. As whole tissue was used to prepare RNA-Seq libraries, mRNA was extracted from the diverse cellular components of the small intestine, including absorptive enterocytes, goblet cells, Paneth cells, enteroendocrine cells, smooth muscle myocytes, leukocytes of the lamina propria, and others. Within the intestine, APOBEC1 is only expressed by the absorptive enterocytes of the luminal epithelium (Funahashi et al., 1995). Therefore, for a given transcript expressed in enterocytes and other intestinal cell types, RNA-Seq reads with mismatches attributable to enterocyte APOBEC1 editing would be "diluted" by non-enterocyte reads. Though not an issue for the enterocyte-specific apoB mRNA (Funahashi et al., 1995), this mixed transcriptome problem could reduce the sensitivity of the screen and could be partially responsible for the apparently low efficiency or incomplete editing observed at newly identified sites. Therefore, in order to more thoroughly investigate APOBEC1 mRNA editing, the comparative RNA-Seq screening approach was applied to small intestinal enterocytes, as described in Chapter 3.

CHAPTER 3: IDENTIFICATION OF APOBEC1 mRNA EDITING TARGETS IN SMALL INTESTINAL ENTEROCYTES

Following the discovery and characterization of APOBEC1 as the catalytic activity responsible for apoB mRNA editing, attempts have been made to identify additional editing targets (Smith et al., 2005). With the exception of NF1 transcript editing in a subset of peripheral nerve sheath tumors (Mukhopadhyay et al., 2002; Skuse et al., 1996), no other physiological mRNA targets have been identified. During the development and validation of a comparative RNA-Seq screening method to detect editing targets (Chapter 2), APOBEC1-specific editing was observed and validated in apoB mRNA as well as several additional transcripts in the small intestine. Although this pilot study was somewhat limited in scale, it prompted a more comprehensive examination of APOBEC1 editing in which the RNA-Seq screening method was applied to small intestinal enterocytes. With additional read depth and superior tissue preparations, numerous APOBEC1 mRNA editing targets were identified in the screen and subsequently validated by standard techniques. Unlike the well-characterized site in the apoB coding sequence, these newly recognized editing sites are all located in the 3′ UTRs of diverse transcripts. These sites share several characteristic sequence features, including a downstream (3′) motif similar to the mooring sequence in apoB mRNA. Bioinformatics analysis based on these features successfully predicted additional APOBEC1 targets that were not represented in the RNA-Seq screen. Many APOBEC1 editing sites are located within transcript regions significantly conserved in mammalian evolution, implying possible functional importance. These findings dramatically expand

the list of APOBEC1 mRNA editing targets, thereby suggesting additional functions for this enzyme beyond its previously characterized role.

3.1. Experimental Procedures

3.1.1. Mice

All C57/BL6 wild-type and congenic *apobec1*$^{-/-}$ mice were used at 6-8 weeks of age. *Apobec1*$^{-/-}$ mice (Hirano et al., 1996) were generously provided by N. Davidson (Washington University School of Medicine, St Louis, MO).

3.1.2. Isolation of small intestinal enterocytes

Mouse small intestines were removed and washed in Hanks Buffered Saline Solution (HBSS) with Ca^{2+} and Mg^{2+}. Jejunum segments were dissected, everted and cut into ~5 cm pieces. Enterocytes were isolated with a protocol adapted from (Xie et al., 2003): Jejunum segments were washed 5 times in HBSS (supplemented with Ca^{2+} and Mg^{2+}) containing 1% fetal bovine serum (FBS) and then washed once in HBSS (without Ca^{2+} and Mg^{2+}) containing 2% glucose and 2% bovine serum albumin (BSA). Jejunum segments were transferred to enterocyte isolation buffer (HBSS without Ca^{2+} and Mg^{2+}, 1.5 mM EDTA, 0.5 mM DTT) and incubated at 37°C with agitation (120 rpm on rotational shaker) for 30 minutes. Enterocyte cell suspensions were collected, passed through a 70 μm cell strainer, washed and resuspended in TRI Reagent (Ambion) for RNA preparation or processed for immunolabeling and flow cytometry.

3.1.3. Immunolabeling, fluorescence microscopy and flow cytometry

For immunolabeling, enterocyte preparations were pre-incubated with Fc Block (BD Biosciences) and then labeled with PE-Cy7-conjugated antibodies against pan-leukocyte marker CD45 (BD Biosciences). Cells were washed, fixed

and permeabilized with Cytofix/Cytoperm solutions (BD Biosciences). After blocking with 5% goat serum (Invitrogen), enterocyte preps were labeled with polyclonal antibodies against villin-1 (Cell Signaling Technology), an enterocyte-specific marker. Secondary labeling was achieved with AlexaFluor 594-conjugated goat anti-rabbit F(ab')2 fragment.

For flow cytometry, cells were resuspended in acquisition buffer (PBS, 5% FBS) and acquired on an LSR II cell analyzer (BD Biosciences).

For fluorescence microscopy, cells were transferred to slides by Cytospin (Thermo Scientific) centrifugation, labeled with DAPI, and mounted in VECTASHIELD medium (Vector Labs). Images were acquired on an Axioplan 2 fluorescence microscope (Zeiss) and processed with Metamorph software (Molecular Devices).

3.1.4. Preparation of RNA-Seq libraries

Enterocyte total RNA was prepared by TRI Reagent (Ambion) extraction and treated with TURBO DNase (Ambion). RNA-Seq libraries were prepared and sequenced as described in Chapter 2.

3.1.5. Mapping RNA-Seq reads and transcriptome sequence analysis

All read mapping and mismatch analyses were performed as described in Chapter 2.

3.1.6. RNA-Seq transcriptome profiling

Transcriptome expression profiling was performed with the Cufflinks software package (available at: *http://cufflinks.cbcb.umd.edu/*). Wild-type and

apobec1$^{-/-}$ RNA-Seq alignments (generated by TopHat) were mapped to RefSeq gene models and relative transcript abundances were calculated based on read distribution.

3.1.7. RNA-Seq read coverage analysis

RNA-Seq read coverage at single-nucleotide resolution was calculated by merging read pileup statistics (SAMtools) with transcript models of expressed genes. Expressed genes were defined as RPKM ≥ 1.0 by RNA-Seq expression analysis. Expressed genes were subdivided into four groups by quartile (RPKM): Very low, low, moderate, and high expression levels. Genomic coordinates for expressed gene exons were derived from RefSeq transcript annotations. These coordinates were merged with SAMtools pileup output to determine the number of reads covering each nucleotide position in expressed transcripts.

3.1.8. APOBEC1 editing site validation

Genomic DNA and RNA were prepared from wild-type and *apobec1*$^{-/-}$ small intestine enterocytes by standard methods. cDNA was prepared from total RNA by Superscript III reverse transcriptase (Inivitrogen) and random hexamer priming and/or 3'RACE oligo(dT) priming (Invitrogen). Sequences containing potential APOBEC1 editing sites were PCR amplified using TurboPfu high-fidelity polymerase (Stratagene). Primer sequences appear in Supplemental Table S3. Primer extension sequencing was performed by GENEWIZ, Inc (South Plainfield, NJ) using Applied Biosystems BigDye version 3.1. The reactions were then run on Applied Biosystem's 3730xl DNA Analyzer.

For subclone sequencing, PCR products were cloned into pSC-B vectors (Stratagene) and transformation colonies were selected by blue/white screening on X-gal LB agar plates. Individual colonies were picked and sequenced as above.

3.1.9. APOBEC1 editing site features: AU content analysis

Bioinformatic analyses of transcript AU content were performed in collaboration with Michael Mwangi (Rockefeller University, New York, NY).

For mRNA feature analysis, several bioinformatic issues regarding sequence annotations were addressed. As the RNA-Seq screen operates on genomic coordinates, analysis considerations were made to adjust for editing at the transcript level. Due to alternative splicing, a gene can generate multiple mRNA isoforms, and when the transcribed but untranslated 3' sequence of a gene contains multiple exons, multiple 3' UTR isoforms may exist. As a result, a single APOBEC1 edit site can appear in different mRNA transcript and 3' UTR annotations. Therefore, computations were performed at the DNA level over specific genomic intervals, herein referred to as GIs. These intervals are the exon segments that code for portions of 3' UTR isoforms as defined in the RefSeq database. It is important to emphasize that the term "GI" is used very specifically here and does not simply refer to exons. While there are a total of 26,558 GIs annotated in the mouse genome, there are many more exons.

The issue of overlapping GIs was also considered. It was determined that this issue was a non-factor in computations and therefore was ignored. None of the GIs that contain an APOBEC1 edit site overlap with another GI, and <5% of the GIs genome-wide overlap with another GI.

Furthermore, though the computations were performed at the DNA level, the U designation (in place of T) was used because the results relate to an editing event at the transcript level.

In assessing AU enrichment of the GIs that contain editing sites, the AU content of a set of sequences was defined simply as the number of A- and U-nucleotides divided by the number of total nucleotides in all sequences of the set. For comparison, random sets of GIs were constructed as follows:

a. For each edit-site-containing GI, a GI was randomly selected from a pool of GIs of comparable length (±20%) according to a uniform distribution over the pool. If the length of the edit-site-containing GI is l, then the length of the new GI was ensured to be in the interval $[0.8l, 1.2l]$. Since the GIs are of varying sizes, it was important to control for length. For example, very long GIs may have sparsely distributed functional elements and so may have long stretches that are not subject to purifying selection. Depending on the edit-site-containing GI, there were a minimum of 262 GIs of comparable length to randomly choose from and a maximum of 1037, with a median of 949. Therefore, there was ample choice, so the random sets were diverse and rarely contained the same elements.
b. A random set was allowed to include edit-site-containing GIs.
c. A random set was not allowed to contain multiple instances of the same GI. If a GI was randomly selected that was selected before, the repeat instance was discarded, and the random selection was redone until a unique GI was obtained.

The AU content of the edit-site containing GIs was compared to the AU content of each of 100,000 comparable random sets of GIs to evaluate significant enrichment.

In assessing the AU-richness of 101 nt windows centered on the edit sites, the AU content of a set of sequences is defined as above. For cases in which multiple edit sites occurred within the same 101 nt window, sites were condensed to a single coordinate. For statistical comparisons, random sets of 101 nt windows were constructed as follows:

a. For each edit-site-containing window, a window of the same length was randomly selected from within the same GI according to a uniform distribution over the GI. The GIs had lengths ranging from 299-4629 nt, with a median of 1380 nt. Therefore, there was ample choice, so the random sets were diverse and rarely contained the same elements.
b. A random set was allowed to include windows that overlapped with edit sites.
c. A random set was not allowed to contain windows that overlapped with each other – overlapping windows result in double counting and can arise in the case of edit sites in the same GI. If a window was randomly selected that overlapped with a previously selected window, the instance was discarded, and the random selection was redone until a non-overlapping window was obtained.

The AU content was computed for the edit site set and compared to the AU content of each of 100,000 comparable random sets to evaluate significant enrichment.

3.1.10. APOBEC1 editing site features: Adjacent nucleotide analysis

Following sequence alignment by editing sites, the bionomial test was used to assess an apparent preference for A and U nucleotides at positions immediately flanking the target cytidine. For N editing sites, the total number of nucleotides in a given sequence alignment column is always equal to N. Let k be the number of A- and U-nucleotides in the column. P-values for each column were computed from the binomial distribution:

$$P = \sum_{i=k}^{N} \binom{N}{i} (f_{\mathrm{AU}})^{i} (1 - f_{\mathrm{AU}})^{N-i}$$

The background AU frequency f_{AU} was calculated as described above. The reported P-values represent the probability of observing $\geq k$ A- or U-nucleotides in a given column under the null hypothesis that A- or U-nucleotides occur with the background frequency f_{AU}.

3.1.11. APOBEC1 editing site features: Sequence motif analysis

Sequence motif analysis was performed with the Multiple Em for Motif Elicitation (MEME) algorithm (*http://meme.nbcr.net/meme4_3_0/cgi-bin/meme.cgi*) (Bailey and Elkan, 1994) on a set 101 nt sequences centered on the editing sites. Statistical significance was approximated by comparing the log likelihood ratio and E-value of the best reported hit to those of the top hit returned by an identical analysis of randomly-shuffled input sequence.

All logo and frequency plots were generated with WebLogo (*http://weblogo.berkeley.edu/*) (Crooks et al., 2004).

3.1.12. APOBEC1 sequence pattern analysis

SequenceSearcher software (*http://athena.bioc.uvic.ca/tools/SequenceSearcher*) (Marass and Upton, 2009) was used to perform regular expression searches for the APOBEC1 consensus sequence pattern within a compiled collection of all RefSeq exon sequences. The list of predicted sites was filtered on gene expression level in wild-type small intestinal enterocytes. When sufficient coverage was available (≥ 3 reads), wild-type RNA-Seq reads mapped to each site were examined for evidence of editing (C:T mismatches above a background frequency of 0.075).

3.1.13. Assessment of phylogenetic conservation

Phylogenetic analyses of conservation and C-T mutation bias were performed in collaboration with Michael Mwangi (Rockefeller University, New York, NY).

PhastCons scores were used to evaluate phylogenetic conservation. In the phastCons analysis, a score is assigned to each nucleotide of the mouse genome in reference to a multialignment of 19 other genomes (placental mammals). The score is in the interval [0, 1] and reflects the degree of conservation across the included species genomes (Siepel et al., 2005). As the mouse serves as the "reference" species in this analysis, scores are not assigned to insertions or deletions in the mouse genome relative to the multialignment.

For a set of 101 nt windows in mouse GIs, the mean phastCons score was computed as follows. First, the sum of the phastCons scores over all nucleotides in all of the windows was computed. This cumulative score was then divided by the total number of nucleotides within the set of windows. Random sets of

windows comparable to those containing APOBEC1 editing sites were constructed as described above. The mean phastCons score was computed for each of 10,000 random sets of windows and compared to the mean phastCons score of the editing site set, with the corresponding P-value indicating the significance of phylogenetic conservation.

3.1.14. Assessing C-T bias at APOBEC1 sites in genomic multi-alignments

An observed bias of C and T nucleotides in multispecies genomic alignments to mouse edit sites was examined quantitatively by statistical comparison. For a site *X* in the placental mammalian genomic alignment, let D(*X*) equal the number of times the C (G) in mouse differs from a base in another mammal, not including indels. Let d(*X*) equal the number of times the C (G) in mouse is replaced by a T (A) in another mammal. The following quantity was used as a measure of the C-to-T (G-to-A) bias for validated APOBEC1 editing sites:

$$B = \left[\sum_{\text{over the 32 sites}} \mathrm{d}(X)\right] \Big/ \left[\sum_{\text{over the 32 sites}} \mathrm{D}(X)\right].$$

Essentially, *B* is the fraction of changes that involve C-to-T (G-to-A) differences between mouse and the other mammals. The larger *B* is, the higher the bias. The value of *B* for the APOBEC1 editing sites was compared to values for random sites, and the results were used to compute a P-value.

Though APOBEC1 editing at the mRNA level always occurs at cytidine residues, editing sites within (-)-strand transcripts are represented as G in genomic context. For analyzing nucleotide bias within genomic multi-alignments, it is important to take strand information into account, as strand

asymmetries may exist with regard to phylogenetic substitutions. For example, it has been reported that C-to-T substitutions do not occur at equal rates on both strands, which is equivalent to the statement that C-to-T and G-to-A substitutions occur at different rates on the (+) strand (Green et al., 2003). Therefore, only (+) strand bases in the mouse genome were considered in the multi-alignment analysis presented here.

An example of a random set of multi-species alignment appears in Figure 3.1B. The random alignment sites used to assess the significance of B were subject to several important constraints. For each edit site-containing row X, a set R(X) of rows was considered. For X, R(X) would serve as the pool from which rows would be randomly drawn. The set R(X) contained any row Y in the multi-species alignment that satisfied the following conditions:

a. Y must be within a 3′ UTR GI.
b. The base in mouse in Y must be identical to the base in mouse in X (e.g. both rows 1 in Figures 3.1 have a G).
c. Y must have the same number of indels as X (e.g. both rows 1 have 7 indels).
d. D(Y) = D(X), where the quantity D(X) was defined above (e.g. both rows 1 have a value of 10).
e. The phastCons score for Y must agree with the phastCons score for X to within ±0.2. This constraint is almost redundant. A row Y that satisfies criterions (c)-(d) often satisfies this criterion.

Hence, the pool R(X) consisted of rows that are very similar to X – they have the same base composition in mouse and nearly identical degrees of conservation.

Depending on X, R(X) contained a minimum of 530 elements and a maximum of 44,023, with a median of 13,449.5. Therefore, there was ample choice, so the random sets were diverse and rarely contained the same elements.

3.1.15 Estimation of miRNA target sites

In order to estimate if APOBEC1 editing affects miRNA targeting, two sets of sequences were assembled: one set in which 13 nt sequences were centered on the edited cytidines (6 nt upstream, 6 nt downstream) and one set in which 13 nt sequences were centered on the editing sites as uridine (6 nt upstream, 6 nt downstream). These sequences were queried against known mature miRNAs (miRBase, *http://www.mirbase.org/*) to determine if any 7 nt substrings is a known miRNA seed.

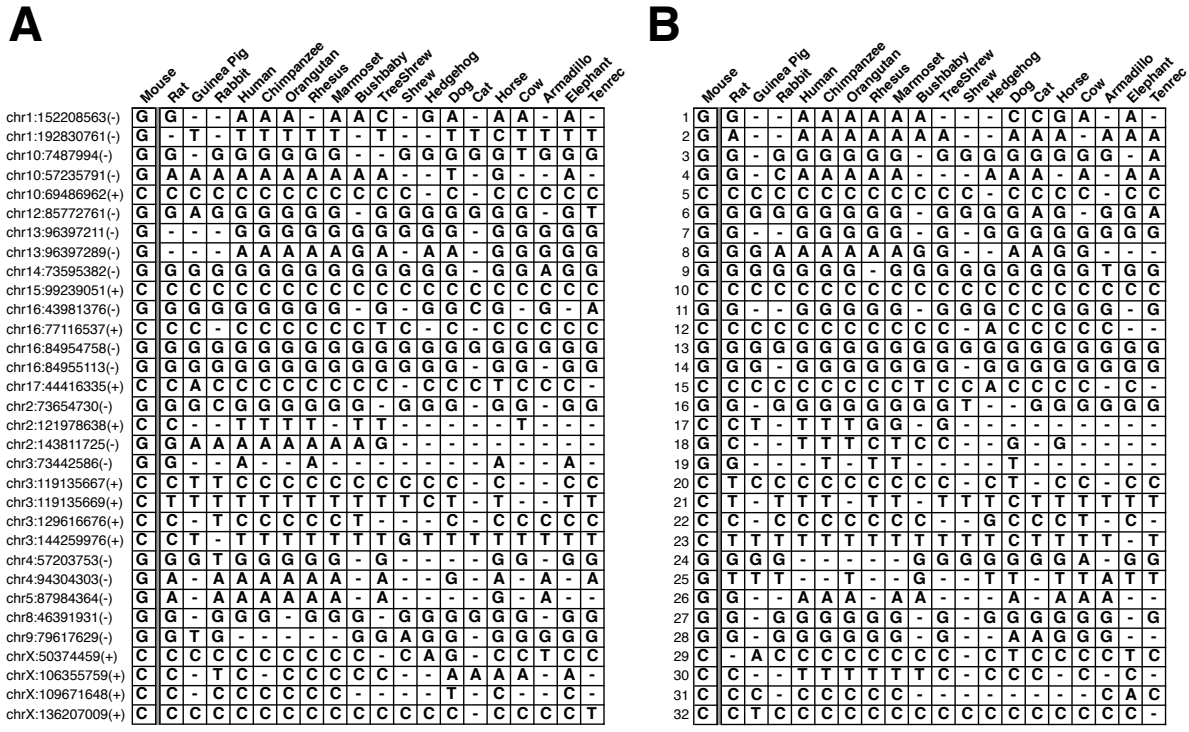

A
Mouse
Rat
Guinea Pig
Rabbit
Human
Chimpanzee
Orangutan
Rhesus
Marmoset
Bushbaby
TreeShrew
Shrew
Hedgehog
Dog
Cat
Horse
Cow
Armadillo
Elephant
Tenrec
chr1:152208563(-) G G - - A A A - A A C - G A - A A - A -
chr1:192830761(-) G - T - T T T T T - T - - T T C T T T T
chr10:7487994(-) G G - G G G G G G - - G G G G G T G G G
chr10:57235791(-) G A A A A A A A A A A - - T - G - - A -
chr10:69486962(+) C C C C C C C C C C C C - C - C C C C C
chr12:85772761(-) G G A G G G G G G - G G G G G G G - G T
chr13:96397211(-) G - - - G G G G G G G G G - G G G G G
chr13:96397289(-) G - - - A A A A A G A - A A - G G G G G
chr14:73595382(-) G G G G G G G G G G G G G G - G G A G G
chr15:99239051(+) C C C C C C C C C C C C C C C C C C C C
chr16:43981376(-) G G G G G G G G G - G - G G C G - G - A
chr16:77116537(+) C C C - C C C C C C T C - C - C C C C C
chr16:84954758(-) G G G G G G G G G G G G G G G G G G G G
chr16:84955113(-) G G G G G G G G G G G G G G - G G - G G
chr17:44416335(+) C C A C C C C C C C C - C C C T C C C -
chr2:73654730(-) G G G C G G G G G G - G G G - G G - G G
chr2:121978638(+) C C - - T T T T - T T - - - - - T - - -
chr2:143811725(-) G G A A A A A A A A G - - - - - - - - -
chr3:73442586(-) G G - - A - - A - - - - - - A - - A -
chr3:119135667(+) C C T T C C C C C C C C C C - C - - C C
chr3:119135669(+) C T T T T T T T T T T C T - T - - T T
chr3:129616676(+) C C - T C C C C C T - - - C - C C C C C
chr3:144259976(+) C C T - T T T T T T G T T T T T T T T
chr4:57203753(-) G G G T G G G G G - G - - - - G G - G G
chr4:94304303(-) G A - A A A A A A - A - - G - A - A - A
chr5:87984364(-) G A - A A A A A A - A - - - - G - A - -
chr8:46391931(-) G G - G G G - G G G - G G G G G G - G G
chr9:79617629(-) G G T G - - - - - G G A G G - G G G G G
chrX:50374459(+) C C C C C C C C C C - C A G - C C T C C
chrX:106355759(+) C C - T C - C C C C C - - A A A A - A -
chrX:109671648(+) C C - C C C C C C - - - - T - C - - C -
chrX:136207009(+) C C C C C C C C C C C C C C - C C C C T
B
Mouse
Rat
Guinea Pig
Rabbit
Human
Chimpanzee
Orangutan
Rhesus
Marmoset
Bushbaby
TreeShrew
Shrew
Hedgehog
Dog
Cat
Horse
Cow
Armadillo
Elephant
Tenrec
1 G G - - A A A A A A - - - C C G A - A -
2 G A - - A A A A A A A - - A A A - A A A
3 G G - G G G G G G - G G G G G G G - A
4 G G - C A A A A A - - - A A A - A - A A
5 C C C C C C C C C C C C - C C C C - C C
6 G G G G G G G G G - G G G A G - G G A
7 G G - - G G G G G - G - G G G G G G G
8 G G G A A A A A A G G - - A A G G - - -
9 G G G G G G G - G G G G G G G G T G G
10 C C C C C C C C C C C C C C C C C C C
11 G G - - G G G G G - G G G C C G G G - G
12 C C C C C C C C C C C C - A C C C C C - -
13 G G G G G G G G G G G G G G G G G G G G
14 G G G - G G G G G G G - G G G G G G G G
15 C C C C C C C C C T C C A C C C C - C -
16 G G - G G G G G G G T - - G G G G G G G
17 C C T - T T T G G - G - - - - - - - - -
18 G C - - T T T C T C C - - G - G - - - -
19 G G - - - T - T T - - - - T - - - - - -
20 C T C C C C C C C C C - C T - C C - C C
21 C T - T T T - T T - T T T C T T T T T T
22 C C - C C C C C C C - - G C C C T - C -
23 C T T T T T T T T T T T C T T T T - T
24 G G G G - - - - - G G G G G G G A - G G
25 G T T T - - T - - G - - T T - T T A T T
26 G G - - A A A - A A - - - A - A A A - -
27 G G - G G G G G G - G - G G G G G G - G
28 G G - G G G G G G - G - - A A G G G - -
29 C - A C C C C C C C C - C T C C C C T C
30 C C - - T T T T T T C - C C C - C - C -
31 C C C - C C C C C - - - - - - - C A C
32 C C T C C C C C C C C C C C C C C C C -

Figure 3.1. Genome multi-alignments for assessment of C-T bias. (A) Multi-alignments for APOBEC1 editing sites, (+) strand context. (B) Example of multi-alignments for random sites.

3.2. Results

3.2.1. RNA-Seq screen for APOBEC1 mRNA editing targets in enterocytes

In order to more thoroughly evaluate APOBEC1 mRNA editing, the comparative RNA-Seq screen was applied to small intestinal enterocytes. Jejunal enterocytes were isolated from the small intestine of C57/BL6 wild-type and congenic *apobec1*$^{-/-}$ mice (Figure 3.2 and Figure 3.3). RNA-Seq libraries were prepared from poly-A^{+} mRNA and deep sequenced, generating 76,766,760 (wild-type) and 50,509,000 (*apobec1*$^{-/-}$) 36 nt reads. Reads were trimmed and aligned to the mouse reference genome (mm9, NCBI 37.1) using the mapping strategy described in Chapter 2 (up to 2 mismatches per sequence, quality conscious, keeping only uniquely mappable reads). Satisfactory alignments were achieved for 42,770,803 and 28,877,750 reads for wild-type and *apobec1*$^{-/-}$, respectively (Table 3.1). Read coverage of transcripts at single base resolution was extensive, particularly for genes expressed at moderate to high levels (Figure 3.4).

Following RNA-Seq read mapping, candidate APOBEC1 mRNA editing sites were identified using the analysis strategy described in Chapter 2. Briefly, a modified SNP-calling algorithm was used to find those sites within RefGene exons at which the reference genome contained a C and wild-type reads contained Ts (or reference G and RNA-Seq read As for (-)-strand transcripts, in genomic context). These sites were then compared to *apobec1*$^{-/-}$ reads. If the corresponding position in *apobec1*$^{-/-}$ reads also contained the mismatch, the site was discarded as a likely genomic polymorphism or non-APOBEC1 modification. However, if the corresponding location in *apobec1*$^{-/-}$ reads matched

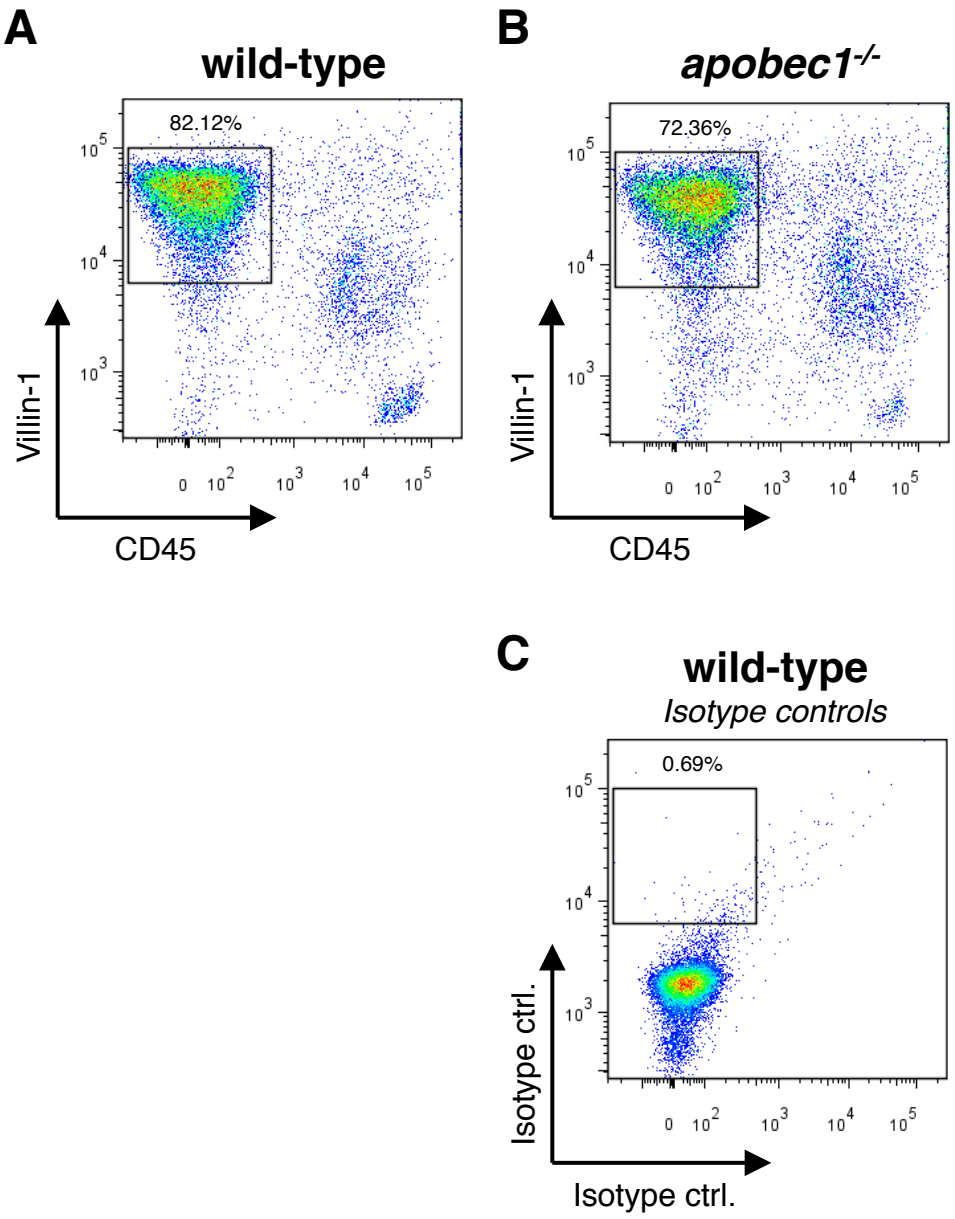


Figure 3.2. Small intestinal enterocytes for RNA-Seq library preparations (flow cytometry). Enterocytes were isolated from wild-type and *apobec1*$^{-/-}$ mice, labeled with villin-1 and CD45 antibodies, and analyzed by flow cytometry. Villin-1^{+} CD45^{-} enterocytes were the predominant cell population, with small numbers of contaminating Villin-1^{-} CD45^{+} leukocytes. (A) wild-type, (B) *apobec1*$^{-/-}$, and (C) wild-type cell preparation labeled with isotype control antibodies.

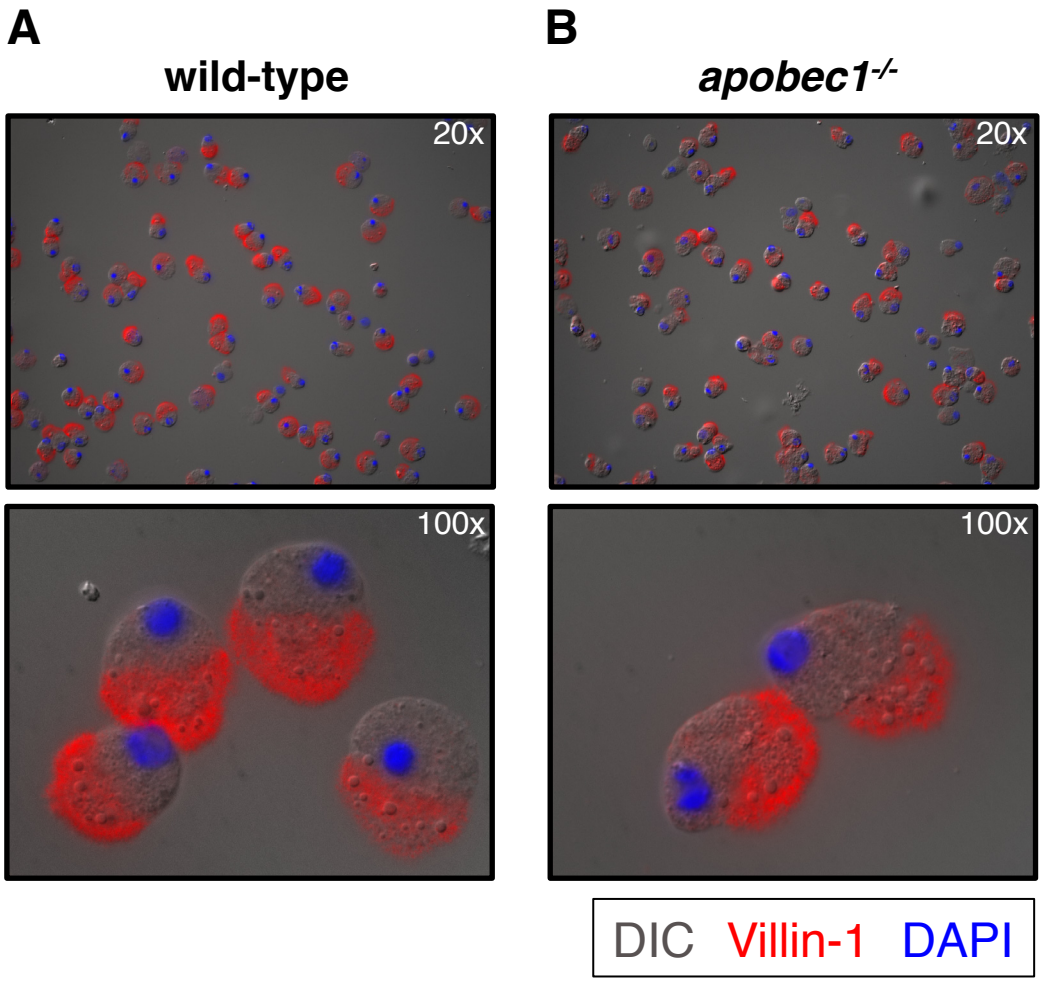


Figure 3.3. Small intestinal enterocytes for RNA-Seq library preparations (immunofluorescence). Enterocytes were isolated from wild-type and *apobec1*$^{-/-}$ mice, labeled with villin-1 and CD45 antibodies, and analyzed by Cytospin fluorescence microscropy. Villin-1 labeling illustrates a polarized organization consistent with intestinal epithelial structure. At high magnification (100x), characteristic microvilli are visible. (A) wild-type, and (B) *apobec1*$^{-/-}$.

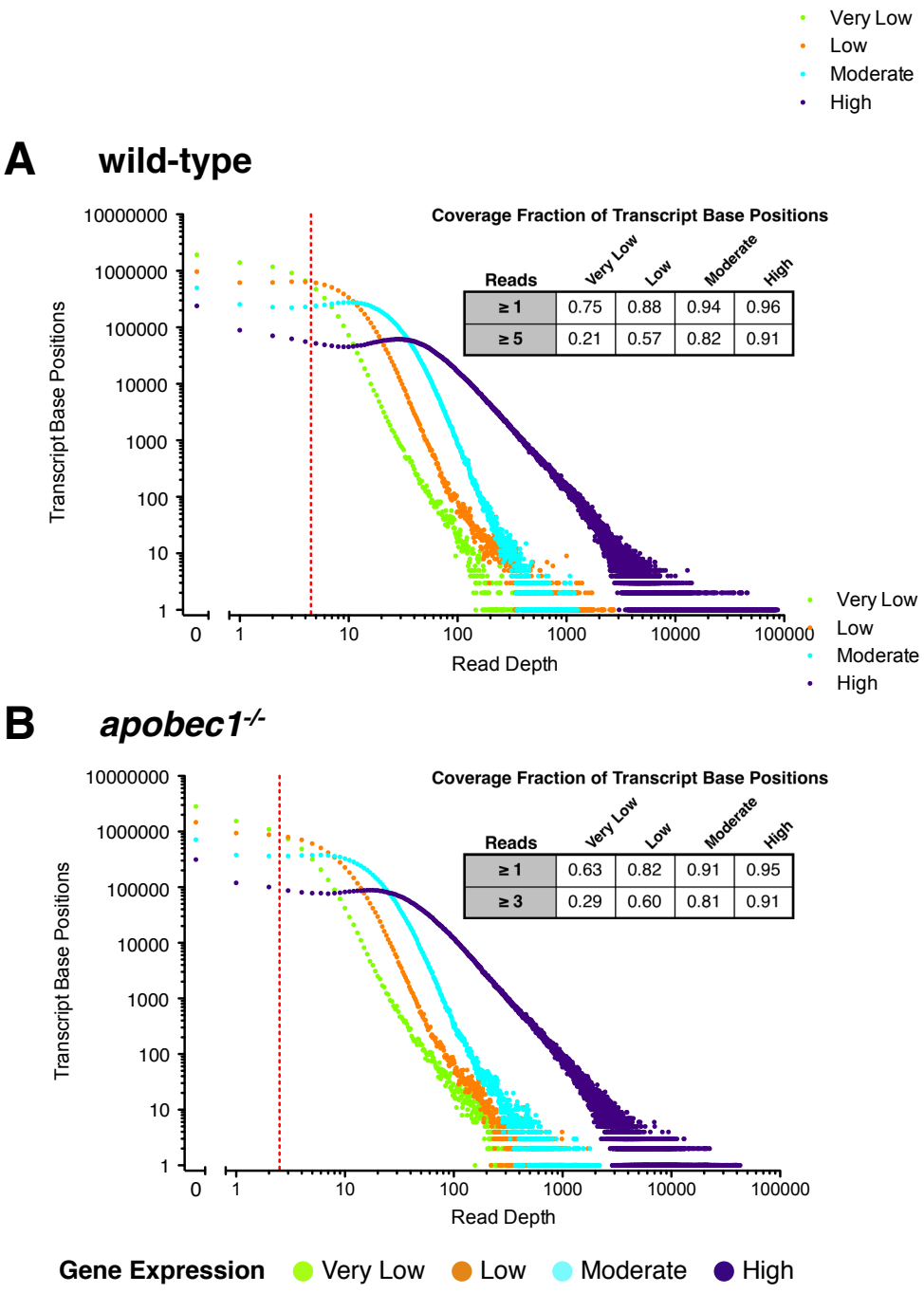


Figure 3.4. Estimation of RNA-Seq transcript coverage. Genes expressed in small intestine enterocytes were divided into expression groups (very low, low, moderate, high) by quartile. Plots represent the number of individual base positions of expressed transcripts covered by the indicated number of mapped RNA-Seq reads. Dashed red line indicates the cutoff for inclusion in candidate editing site analyses. Inset tables report the fraction of individual base positions covered by at least 1 RNA-Seq read and at least the candidate editing site analysis cutoff value of mapped RNA-Seq reads. (A) wild-type, and (B) *apobec1*$^{-/-}$.!

Table 3.1. RNA-Seq read dataset statistics

	Raw Reads	Uniquely Mapped Reads
wild-type	76,766,760	42,770,803 (*56%*)
***apobec1*$^{-/-}$**	50,509,000	28,877,750 (*57%*)

Table 3.2. Candidate APOBEC1 editing site statistics by analysis filter

Analysis Filter	Candidate edit sites remaining
wild-type read:reference mismatches (*unfiltered*)	44,250
Retain sites mapped to RefSeq exons	1,716
Retain reference **C** / read **T** mismatches	194
Remove known SNPs	181
Retain APOBEC1-specific mismatches (no mismatch in *apobec1*$^{-/-}$ read set)	93
Remove low read depth / low confidence sites	43
Remove mapping artifacts	39
Validate by Sanger sequencing	33

the reference sequence, the site was considered for additional analysis . After filtering out those sites with insufficient read coverage (<5 reads for wild-type, <3 for *apobec1*$^{-/-}$) and/or mismatch probability scores (Table 3.2), 39 remaining sites were designated candidate APOBEC1 mRNA editing targets. As in the whole intestine pilot study, when these sites were ranked by mismatch probability score, the top hit was the apoB mRNA editing target. Once again, unbiased detection of this positive control site indicated a successful screen and supported further analysis of the additional candidate targets. An example of RNA-Seq read alignments is presented in Figure 3.5.

3.2.2. Validation of candidate editing sites in enterocytes

To validate the potential editing events identified in the RNA-Seq screen, standard dideoxynucleotide Sanger sequencing was used to examine the sites in genomic DNA and RNA (cDNA) isolated from intestinal enterocytes. All validation samples were independently prepared from different mice than those used for RNA-Seq libraries. Sanger sequencing results for several sites are presented in Figure 3.6. Clear evidence of C-to-U(T) RNA editing was observed at 33 of the 39 candidate sites. Sequencing results for 6 sites did not indicate editing; these candidates were rejected as false positive events. The remaining C/T chromatogram peaks in wild-type cDNA were of varied intensity, indicating differences in editing levels. However, the T chromatogram peaks were considerably more prominent at many sites as compared to sequencing data for whole intestine tissue (Chapter 2 – Figure 2.5). This suggests that edited transcripts in whole intestine preparations were "diluted" by unedited transcripts from non-enterocyte cell types. At one site, chr3:73442586(-),

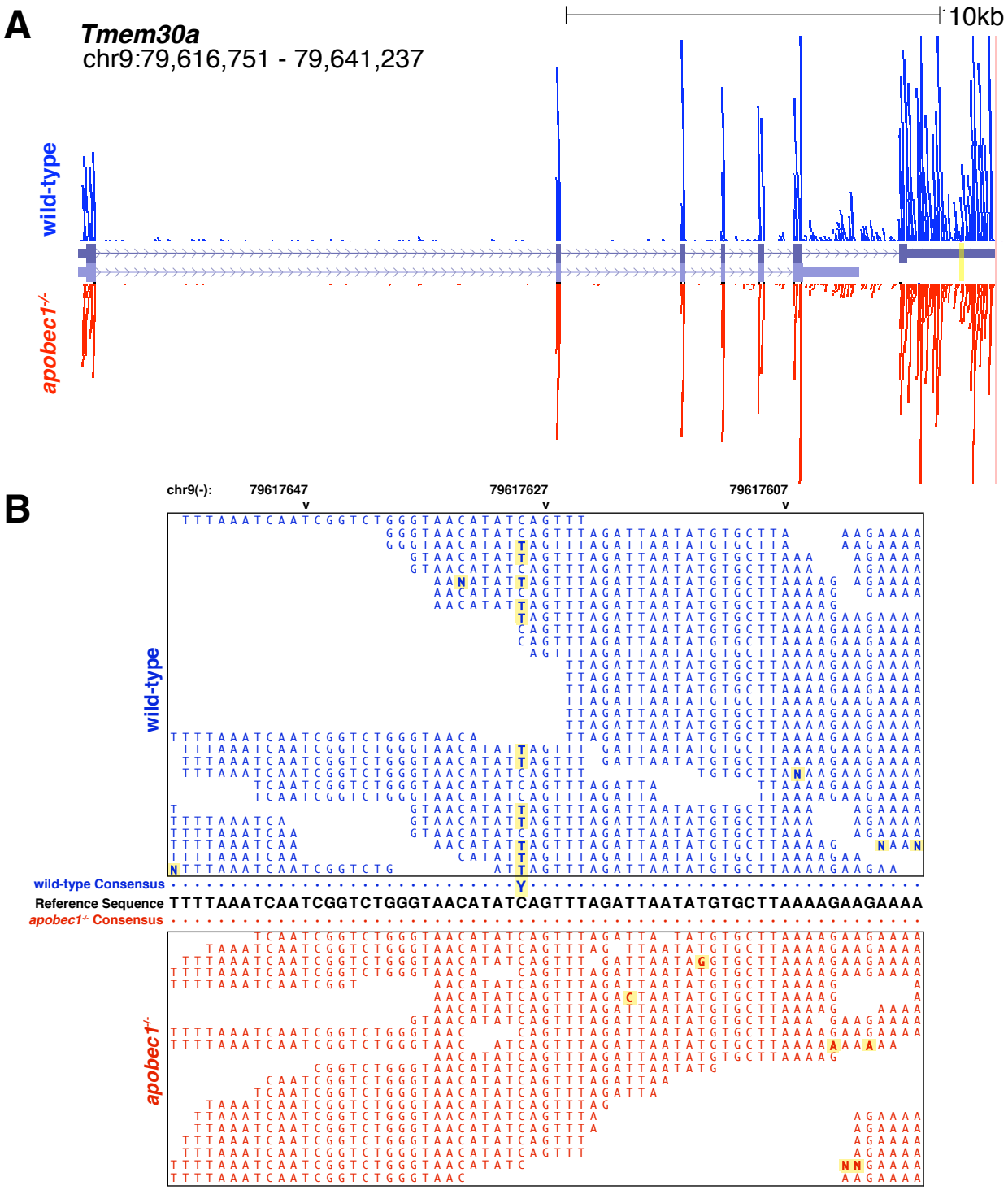


Figure 3.5. Identification of candidate APOBEC1 editing sites by comparative RNA-Seq screen: *Tmem30a*. (A) Genome annotation for *Tmem30a* indicates two alternatively-spliced transcript models. RNA-Seq read (blue squares, wild-type; red squares, *apobec1*$^{-/-}$) distribution suggests that the first (dark blue, upper) is the most abundant isoform in intestinal enterocytes. A potential APOBEC1 edit site was identified in the 3′ UTR. (B) Detail of region containing candidate APOBEC1 edit site (yellow box, *A*). RNA-Seq reads provide overlapping coverage at single-nucleotide resolution. Mismatches to reference are highlighted yellow. The edit site displays numerous T reads in the wild-type sample, but only C reads in the *apobec1*$^{-/-}$ sample.

significant additional editing was observed at a cytidine adjacent to the location identified by the screen. To further validate APOBEC1-specific editing, several sites were selected for subcloning and additional Sanger sequencing (Figure 3.7 and Figure 3.8). C/T mismatches at candidate editing sites were observed only in subclones derived from wild-type cDNA; no deviations from reference sequence were present in wild-type genomic DNA, *apobec1*$^{-/-}$genomic DNA or *apobec1*$^{-/-}$ cDNA. Additionally, low level "hyperediting" of C residues in close proximity to the primary editing site was observed in a minority of subclones for several targets, including apoB. This phenomenon has been previously described for apoB mRNA (Figure 3.9) and is of unknown functional significance (Sowden et al., 1996a; Sowden et al., 1998; Yamanaka et al., 1996).

A list of validated APOBEC1 mRNA editing targets appears in Table 3.3. Unlike the edited coding sequence of apoB mRNA, all of the newly identified APOBEC1 sites are located in transcript 3′ UTRs. RNA-Seq read data were used to estimate the editing level of each site ([# of T reads] / [# of C reads + # of T reads]). ApoB mRNA displayed the most pronounced editing (0.92), with editing frequency of 3′ UTR sites ranging from 0.18 to 0.79 (Figure 3.10). Editing frequencies calculated from RNA-Seq reads were very similar to those determined by cDNA amplification, subcloning, and Sanger sequencing.

3.2.3. Transcriptome profiling of wild-type and *apobec1*$^{-/-}$ enterocytes

Sequences in transcript 3′ UTRs can influence mRNA stability and gene expression by a number of mechanisms, including miRNA targeting, subcellular localization and regulation by RNA binding proteins. Therefore, changes in 3′ UTR sequences by RNA editing could affect transcript and/or protein levels. For

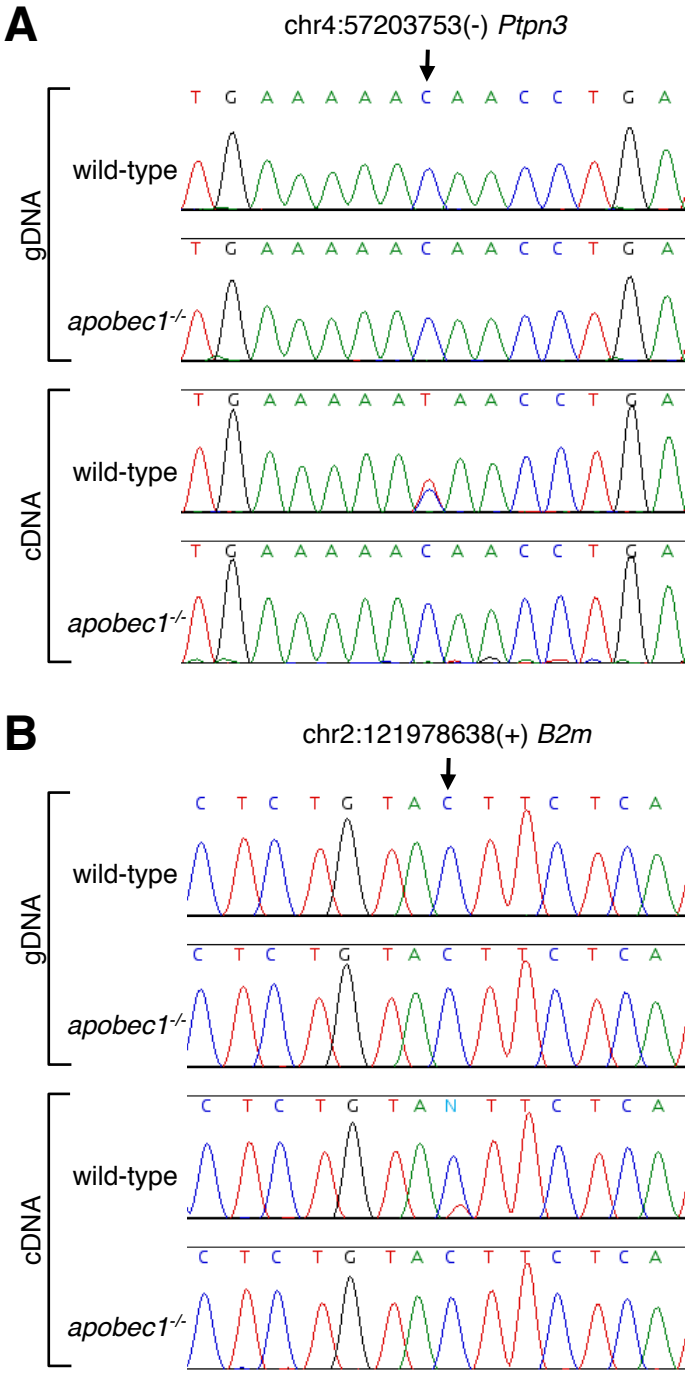


Figure 3.6. Validation of candidate APOBEC1 mRNA editing sites in small intestinal enterocytes. Representative examples of conventional Sanger sequencing chromatograms for wild-type and *apobec1*$^{-/-}$ genomic DNA and cDNA at editing sites are shown. C-to-U(T) editing is apparent only in wild-type cDNA. (A) chr4:57203753(-) in the *Ptpn3* transcript, (B) chr2:121978638(+) in the *B2m* transcript, (C) chr16:43981376(-) in the *Gramd1c* transcript, and (D) chr3:73442586(-) in the *Bche* transcript.

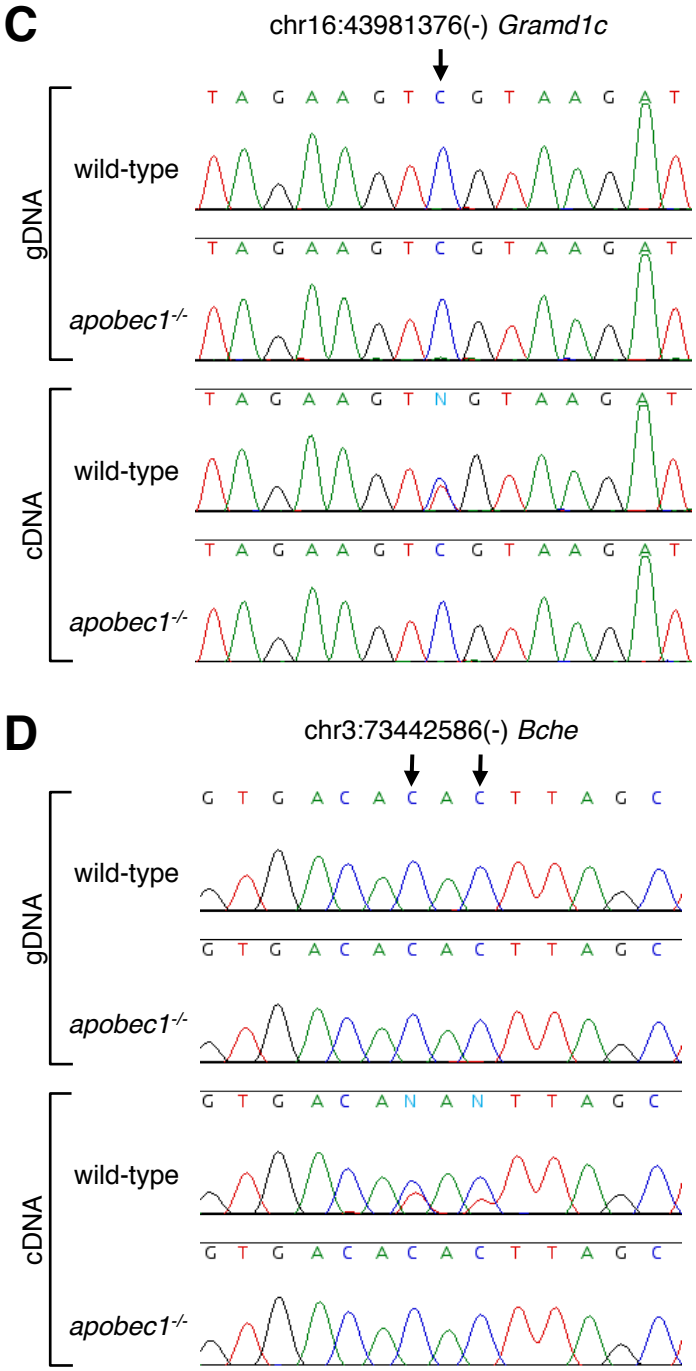


Figure 3.6. Validation of candidate APOBEC1 mRNA editing sites in small intestinal enterocytes, continued.

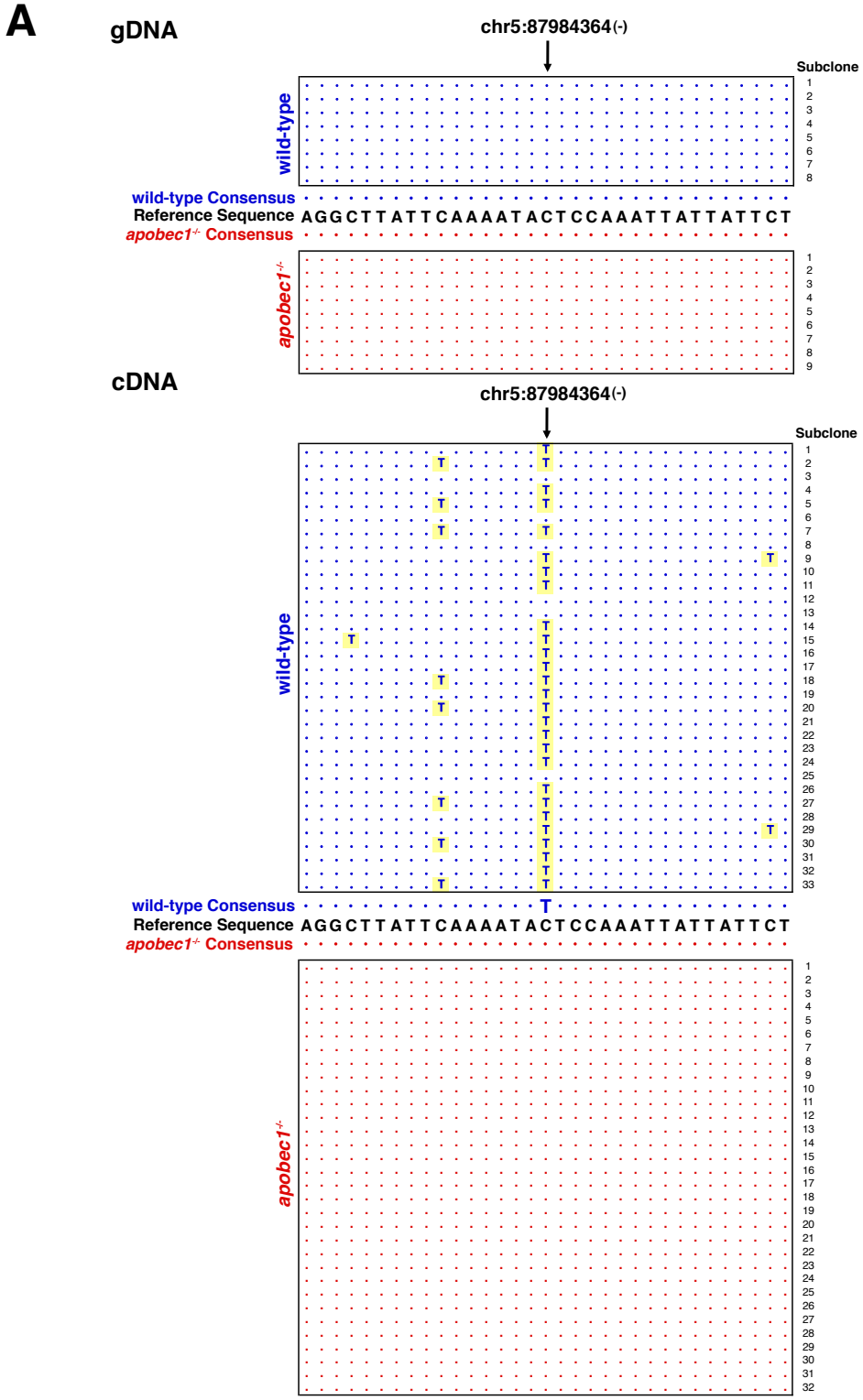


Figure 3.7. Validation of APOBEC1 mRNA edit sites by subclone sequencing. gDNA and cDNA subclones from wild-type and *apobec1*$^{-/-}$ enterocytes were sequenced by conventional Sanger techniques and aligned to transcript reference. Mismatches to reference are highlighted in yellow. Arrows indicate candidate editing sites identified in RNA-Seq screen. (A) chr5:87984364(-) *Sult1d1*.

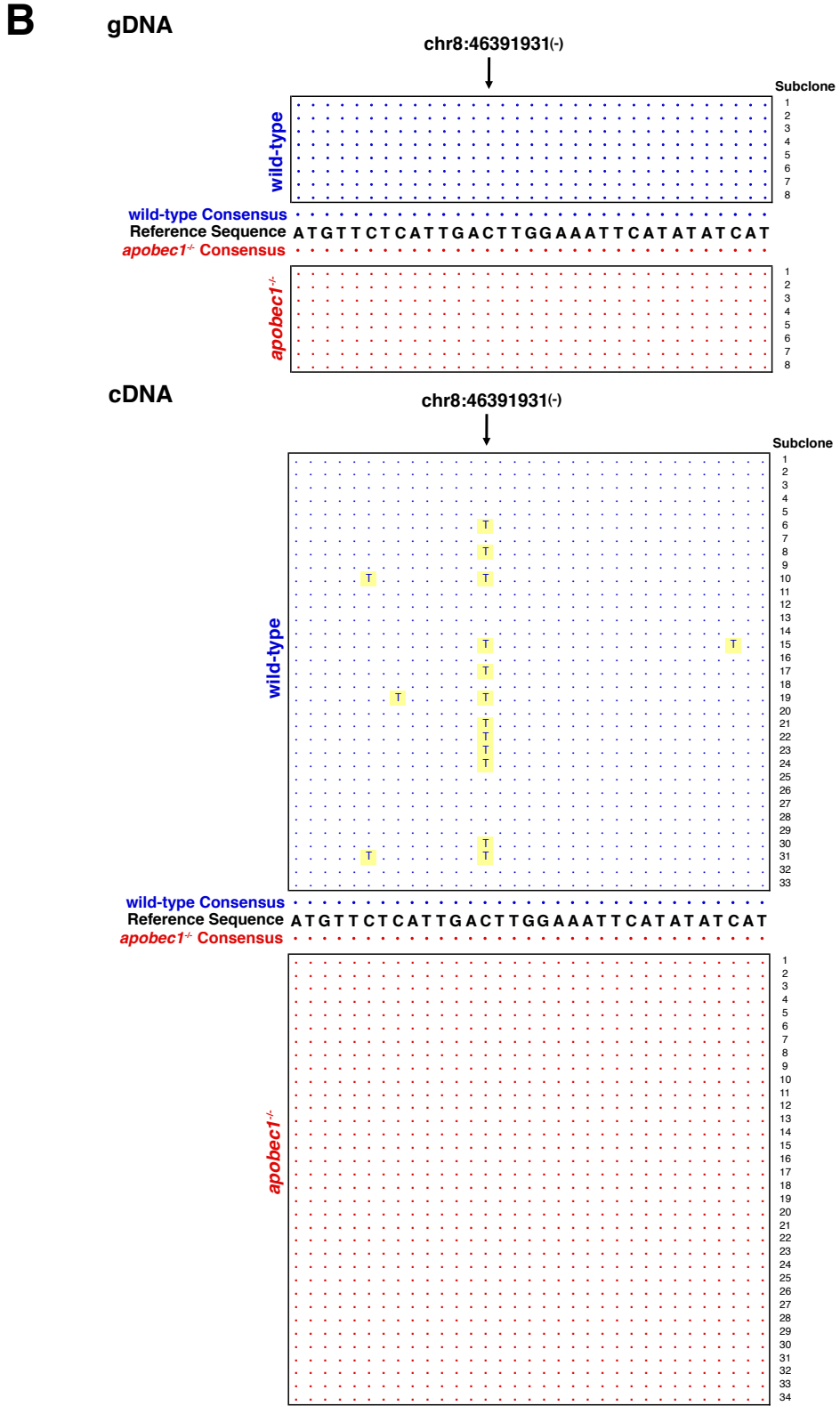


Figure 3.7. Validation of APOBEC1 mRNA edit sites by subclone sequencing, continued. (B) chr8:46391931(-) *Cyp4v3*.

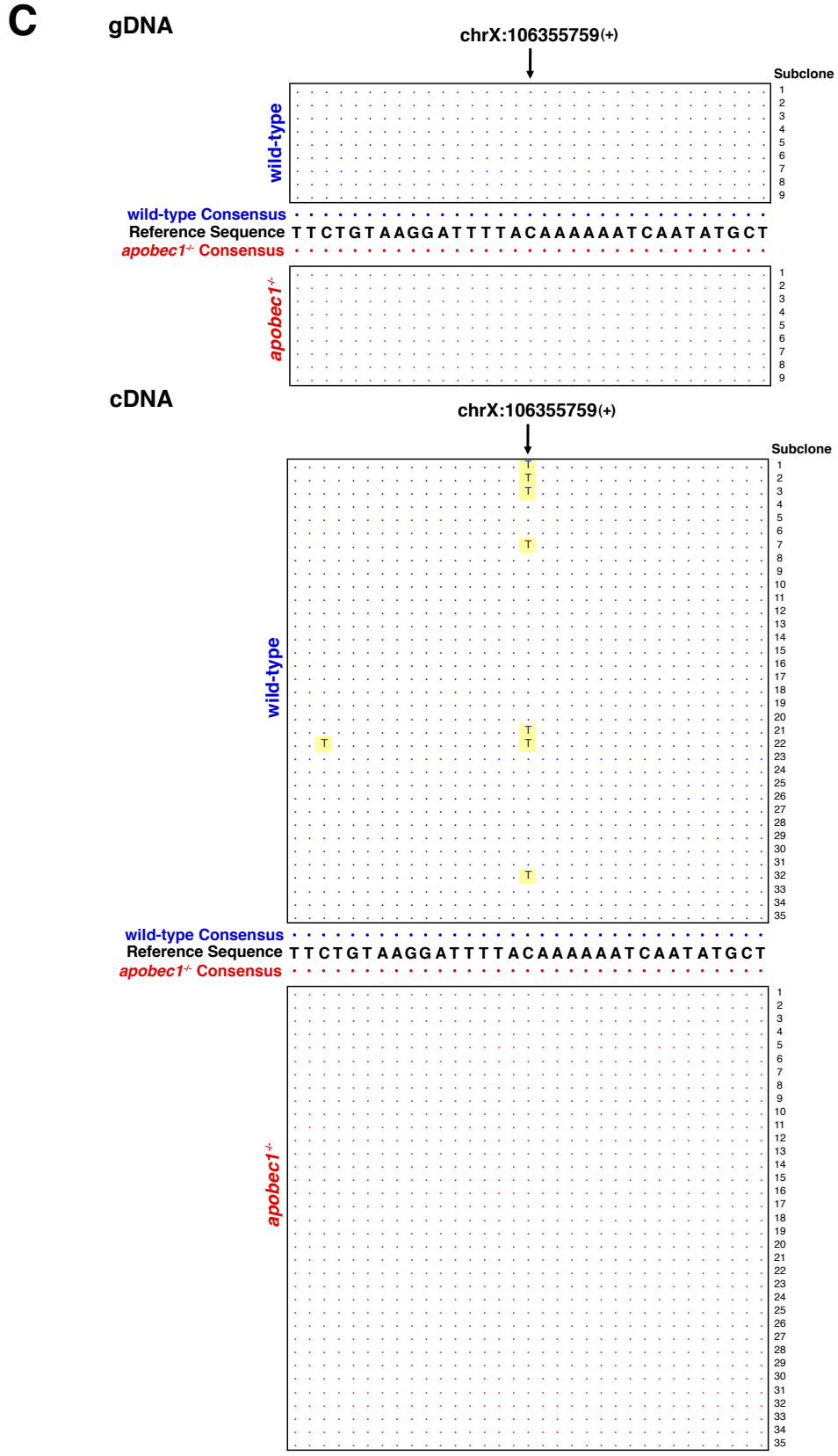


Figure 3.7. Validation of APOBEC1 mRNA edit sites by subclone sequencing, continued. (C) chrX:106355759(+) *Sh3bgrl*.

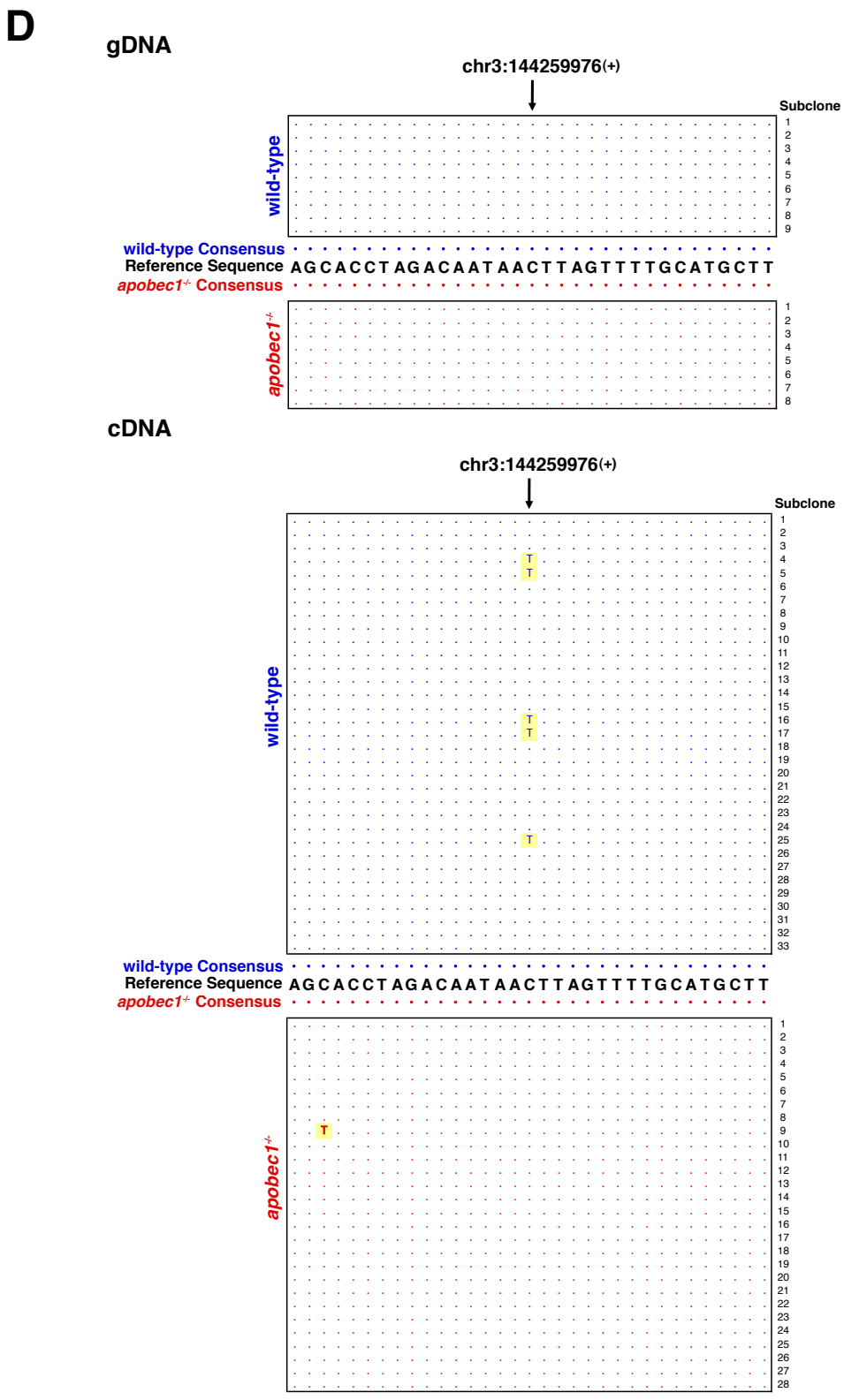


Figure 3.7. Validation of APOBEC1 mRNA edit sites by subclone sequencing, continued. (D) chr3:144259976(+) *Sep15*.

chr5:87984364(-)
NM_016771
Sult1d1

	gDNA		cDNA	
	wild-type	*apobec1*$^{-/-}$	wild-type	*apobec1*$^{-/-}$
A	0	0	0	0
C	8	9	6	32
G	0	0	0	0
T	0	0	27	0
Subclones sequenced	8	9	33	32

chr8:46391931(-)
NM_133969
Cyp4v3

	gDNA		cDNA	
	wild-type	*apobec1*$^{-/-}$	wild-type	*apobec1*$^{-/-}$
A	0	0	0	0
C	8	8	21	34
G	0	0	0	0
T	0	0	12	0
Subclones sequenced	8	8	33	34

chrX:106355759(+)
NM_019989
Sh3bgrl

	gDNA		cDNA	
	wild-type	*apobec1*$^{-/-}$	wild-type	*apobec1*$^{-/-}$
A	0	0	0	0
C	9	9	28	35
G	0	0	0	0
T	0	0	7	0
Subclones sequenced	9	9	35	35

chr3:144259976(+)
NM_053102
Sep15

	gDNA		cDNA	
	wild-type	*apobec1*$^{-/-}$	wild-type	*apobec1*$^{-/-}$
A	0	0	0	0
C	9	8	28	28
G	0	0	0	0
T	0	0	5	0
Subclones sequenced	9	8	33	28

Figure 3.8. Validation of APOBEC1 mRNA edit sites by subclone sequencing. Nucleotide frequencies at candidate APOBEC1 editing sites of subclone sequences presented in Figure 3.7.

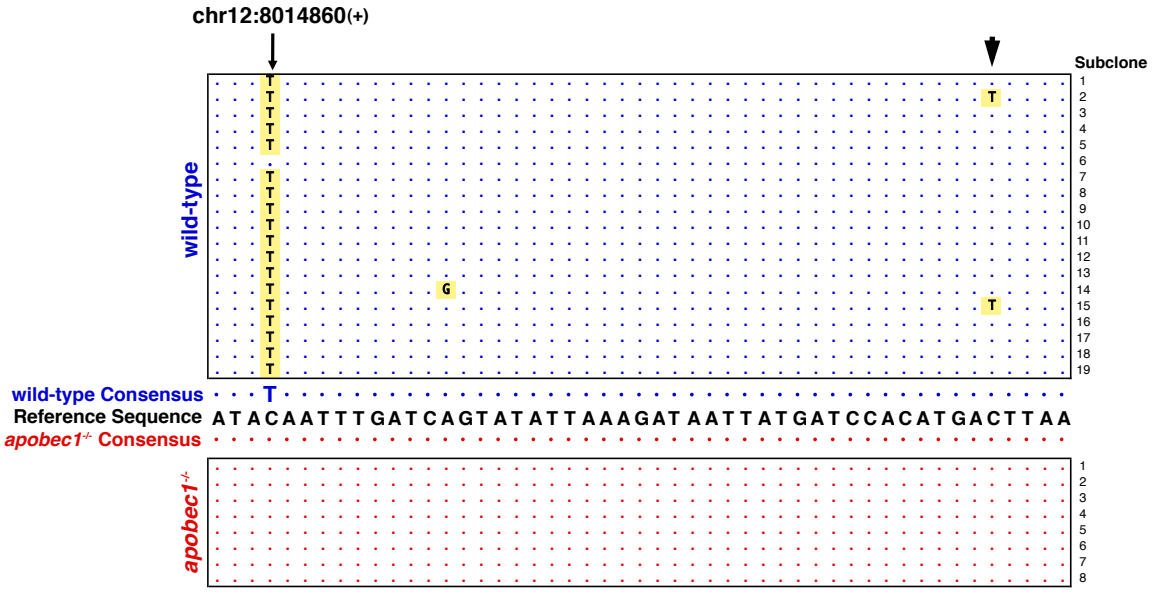


Figure 3.9. Hyperediting of apoB mRNA. Alignments of individual subclone sequences at apoB editing site (indicated by arrow). cDNA subclones from wild-type and *apobec1*$^{-/-}$ enterocytes were sequenced by conventional Sanger techniques and aligned to apoB reference. Hyperediting is apparent in two wild-type subclone sequences (arrowhead).

Table 3.3. Validated APOBEC1 editing sites (small intestinal enterocytes)

				wild-type					*apobec1*$^{-/-}$				
Genome Site	Gene	Type	Ref. Base	Read Cons	P Cons.	P Mism.	Read Depth	Edit Freq.	Read Cons	P Cons.	P Mism.	Read Depth	Edit Freq.
chr12:8014860(+)	Apob	CDS	**C**	**T**	255	255	204	0.93	**C**	255	0	128	0.00
chr2:121978638(+)	B2m	3'UTR	**C**	**Y**	228	228	2860	0.18	**C**	255	0	1582	0.00
chrX:109671648(+)	2010106E10Rik	3'UTR	**C**	**Y**	228	228	688	0.46	**C**	255	0	322	0.00
chr8:46391931(-)	Cyp4v3	3'UTR	**G**	**R**	228	228	112	0.38	**G**	117	0	42	0.00
chr3:129616676(+)	Casp6	3'UTR	**C**	**Y**	228	228	107	0.50	**C**	255	0	119	0.00
chr17:44416335(+)	Clic5	3'UTR	**C**	**Y**	175	175	186	0.31	**C**	255	0	92	0.01
chr10:57235791(-)	Serinc1	3'UTR	**G**	**R**	77	170	29	0.75	**G**	39	0	4	0.00
chr5:87984364(-)	Sult1d1	3'UTR	**G**	**R**	60	154	28	0.79	**G**	65	0	20	0.00
chr2:143811725(-)	Rrbp1	3'UTR	**G**	**R**	149	149	40	0.38	**G**	63	0	23	0.00
chr10:7487994(-)	BC013529	3'UTR	**G**	**R**	141	141	20	0.45	**G**	45	0	6	0.00
chr9:79617629(-)	Tmem30a	3'UTR	**G**	**R**	129	135	22	0.55	**G**	87	0	20	0.00
chr1:152208563(-)	BC003331	3'UTR	**G**	**R**	54	132	23	0.74	**G**	48	0	7	0.00
chr4:57203753(-)	Ptpn3	3'UTR	**G**	**R**	67	124	15	0.67	**G**	48	0	7	0.00
chr16:77116537(+)	Usp25	3'UTR	**C**	**Y**	116	116	16	0.50	**C**	45	0	6	0.00
chr3:119135667(+)	Dpyd	3'UTR	**C**	**Y**	115	115	26	0.32	**C**	63	0	12	0.00
chr16:84955113(-)	App	3'UTR	**G**	**R**	108	108	563	0.21	**G**	255	0	357	0.00
chr13:96397289(-)	Iqgap2	3'UTR	**G**	**R**	103	103	514	0.23	**G**	255	0	387	0.00
chr3:144259976(+)	Sep15	3'UTR	**C**	**Y**	93	103	13	0.54	**C**	42	0	5	0.00
chrX:136207009(+)	Rnf128	3'UTR	**C**	**Y**	91	91	669	0.20	**C**	255	0	397	0.00
chrX:106355759(+)	Sh3bgrl	3'UTR	**C**	**Y**	89	89	23	0.30	**C**	75	0	16	0.00
chrX:50374459(+)	Hprt1	3'UTR	**C**	**Y**	85	85	55	0.22	**C**	108	0	27	0.00
chr4:94304303(-)	Lrrc19	3'UTR	**G**	**R**	85	85	38	0.26	**G**	87	0	20	0.00
chr3:119135669(+)	Dpyd	3'UTR	**C**	**Y**	84	84	25	0.28	**C**	60	0	11	0.00
chr14:73595382(-)	Rb1	3'UTR	**G**	**R**	83	83	21	0.33	**G**	30	0	12	0.00
chr12:85772761(-)	Aldh6a1	3'UTR	**G**	**R**	64	80	9	0.56	**G**	42	0	5	0.00
chr2:73654730(-)	Atf2	3'UTR	**G**	**R**	73	73	21	0.29	**G**	54	0	9	0.00
chr16:43981376(-)	Gramd1c	3'UTR	**G**	**R**	64	64	17	0.29	**G**	51	0	8	0.00
chr16:84954758(-)	App	3'UTR	**G**	**R**	60	60	293	0.21	**G**	255	0	118	0.01
chr10:69486962(+)	Ank3	3'UTR	**C**	**Y**	56	56	11	0.36	**C**	36	0	3	0.00
chr13:96397211(-)	Iqgap2	3'UTR	**G**	**R**	55	55	124	0.38	**G**	150	0	41	0.00
chr3:73442586(-)	Bche	3'UTR	**G**	**R**	54	54	14	0.36	**G**	78	0	17	0.00
chr1:192830761(-)	Mfsd7b	3'UTR	**G**	**A**	2	48	9	0.78	**G**	42	0	5	0.00
chr15:99239051(+)	Tmbim6	3'UTR	**C**	**Y**	45	45	389	0.20	**C**	255	0	196	0.00

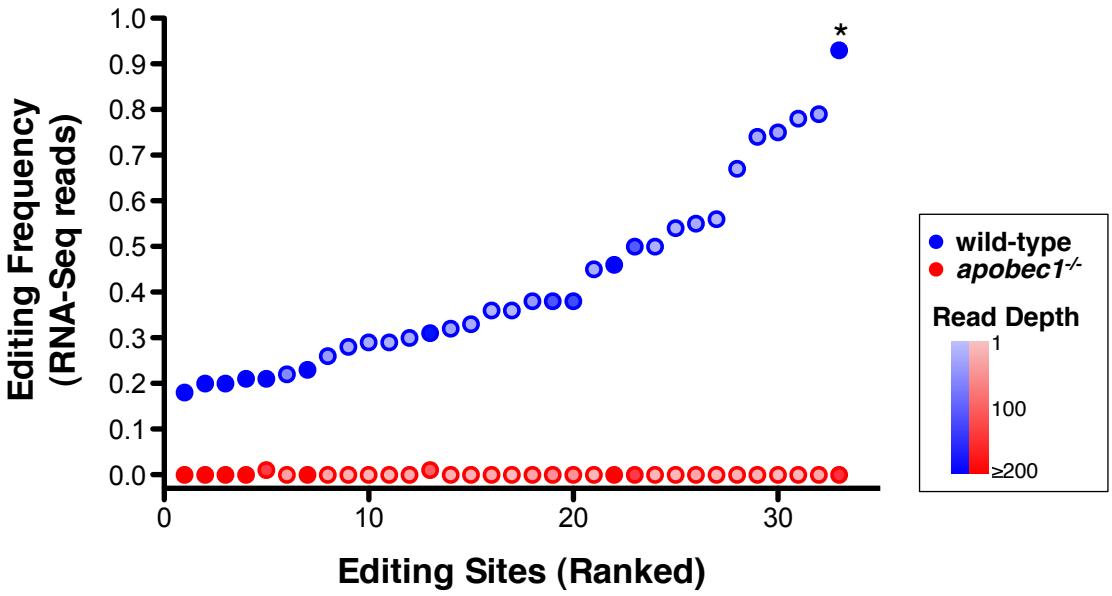


Figure 3.10. Editing frequency at APOBEC1 sites. Sites were ranked by editing frequency, calculated as (# of T reads) / (# of C reads + # of T reads). In contrast to the efficiently edited apoB transcript, sites displayed a broad range of editing frequencies. *apoB editing site.

example, more than 35% of the identified APOBEC1 editing sites are located within sequences that match the seed targets of known miRNAs (Table 3.4). C-to-U nucleotide changes at these sites could disrupt miRNA targeting and correspondingly affect transcript regulation.

Though RNA-Seq was used primarily to screen for sequence mismatches indicative of APOBEC1 editing, the read datasets can also be used for quantifying transcript expression levels. In order to compare transcript levels in wild-type and *apobec1*$^{-/-}$ enterocytes, whole-transcriptome expression profiling was carried out on the RNA-Seq reads generated for the mismatch screen. Comparative analysis indicated that many transcripts were differentially expressed, including a substantial number associated with lipid processing and transport. However, of the transcripts containing 3' UTR APOBEC1 editing sites, most were expressed at similar levels between samples (Figure 3.11). Though a few were observed at different levels, greater than 2-fold alterations were not observed. Therefore, though many transcripts are differentially expressed in wild-type and *apobec1*$^{-/-}$ enterocytes, possibly as a result of apoB-related lipid accumulation, no remarkable differences in expression at the transcript level were observed for APOBEC1 edit site-containing mRNAs.

3.2.4. APOBEC1 mRNA edit sites share characteristic sequence features

Target recognition by RNA editing enzymes is typically determined by the sequence and/or structural context of the edited base (Bass, 2002; Davidson, 2002). Although features contributing to apoB mRNA editing have been previously characterized, it was unclear whether similar attributes would apply to APOBEC1 editing of 3' UTR targets. To determine if the sites identified here

Table 3.4. APOBEC1 editing sites in miRNA seed sequence matches

Editing Site	Transcript Accession	Gene	miRNA Seed Match (C)	miRNA Seed Match (U)
chr1:152208563 (-)	NM_001077237	*BC003331*	mmu-miR-669b	
chr1:192830761 (-)	NM_001081259	*Mfsd7b*		
chr2:73654730 (-)	NM_009715	*Atf2*	mmu-miR-669n	mmu-miR-297a mmu-miR-297b-5p mmu-miR-297c
chr2:121978638 (+)	NM_009735	*B2m*		mmu-miR-539
chr2:143811725 (-)	NM_133626	*Rrbp1*	mmu-miR-539	
chr3:73442586 (-)	NM_009738	*Bche*	mmu-miR-467e mmu-miR-467h mmu-miR-1970 mmu-miR-599	
chr3:119135667 (+)	NM_170778	*Dpyd*		
chr3:119135669 (+)	NM_170778	*Dpyd*		
chr3:129616676 (+)	NM_009811	*Casp6*	mmu-miR-691	
chr3:144259976 (+)	NM_053102	*Sep15*		
chr4:57203753 (-)	NM_011207	*Ptpn3*		mmu-miR-154
chr4:94304303 (-)	NM_175305	*Lrrc19*		
chr5:87984364 (-)	NM_016771	*Sult1d1*	mmu-miR-496	
chr8:46391931 (-)	NM_133969	*Cyp4v3*		
chr9:79617629 (-)	NM_133718	*Tmem30a*	mmu-miR-190 mmu-miR-190b	
chr10:7487994 (-)	NM_145418	*BC013529*		mmu-miR-466l
chr10:57235791 (-)	NM_019760	*Serinc1*		
chr10:69486962 (+)	NM_170729	*Ank3*		
chr12:85772761 (-)	NM_134042	*Aldh6a1*		
chr13:96397211 (-)	NM_027711	*Iqgap2*		
chr13:96397289 (-)	NM_027711	*Iqgap2*	mmu-miR-370 mmu-miR-683	mmu-miR-323-3p
chr14:73595382 (-)	NM_009029	*Rb1*		
chr15:99239051 (+)	NM_026669	*Tmbim6*		
chr16:43981376 (-)	NM_153528	*Gramd1c*	mmu-miR-1964	
chr16:77116537 (+)	NM_013918	*Usp25*		
chr16:84954758 (-)	NM_007471	*App*		
chr16:84955113 (-)	NM_007471	*App*	mmu-miR-186	
chr17:44416335 (+)	NM_172621	*Clic5*	mmu-miR-143	
chrX:50374459 (+)	NM_013556	*Hprt1*		
chrX:106355759 (+)	NM_019989	*Sh3bgrl*		
chrX:109671648 (+)	NM_026333	*2010106E10Rik*	mmu-miR-142-3p	
chrX:136207009 (+)	NM_023270	*Rnf128*		

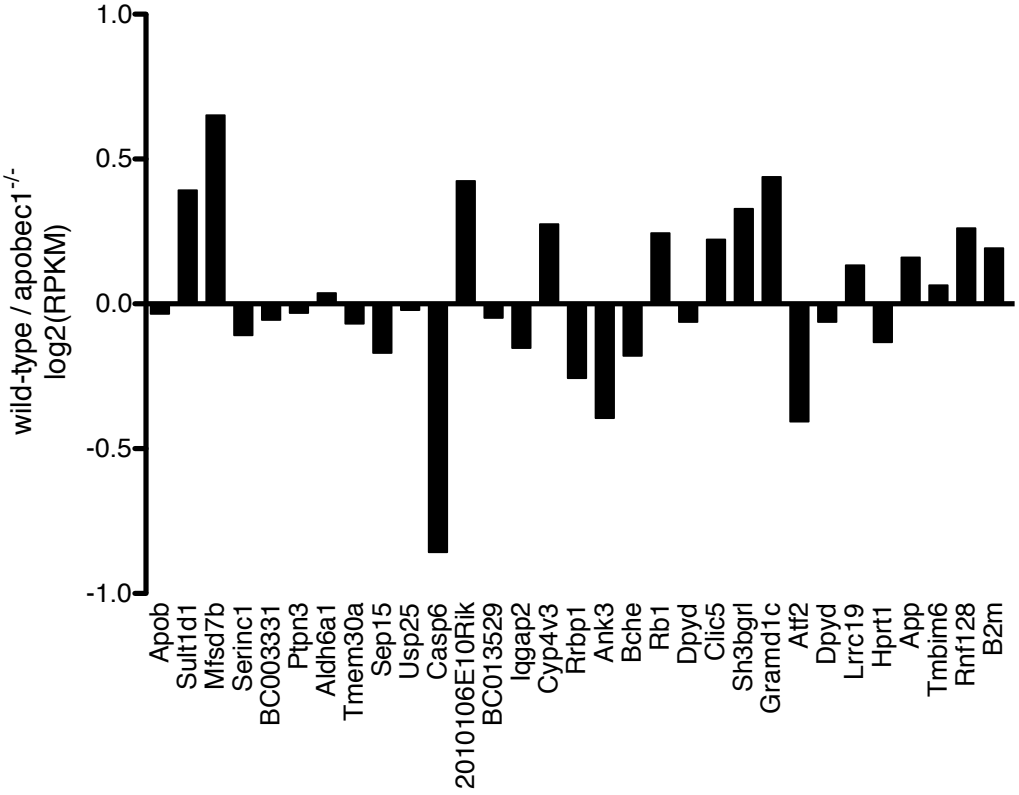


Figure 3.11. Gene expression profiling for APOBEC1 mRNA editing targets. RNA-Seq transcriptome profiling analysis for wild-type and *apobec1*$^{-/-}$ small intestinal enterocytes was performed with the Cufflinks software package. For each gene indicated, plotted values are the $\log_2$ fold-difference (wild-type : *apobec1*$^{-/-}$) mRNA expression in RNA-Seq Reads Per Kilobase of exon model per Million mapped reads (RPKM). Though subtle differences are apparent, dramatic changes (i.e. greater than 2-fold) in mRNA expression were not observed.

share common *cis* features that might "mark" them for APOBEC1 editing, the sequences flanking the edited cytidines were further examined.

APOBEC1 has RNA binding activity with a preference for sequences rich in A and U (Anant et al., 1995; Navaratnam et al., 1995). The sequence region (101 nt) surrounding the apoB mRNA editing site is also particularly AU-rich (0.70 AU content). The AU content of the APOBEC1 edit site 3′ UTRs was computed (0.63) and found to be significantly more AU-rich than comparable sets of 3′ UTRs chosen at random (Figure 3.12A, $p < 0.0001$). Furthermore, within these AU-rich 3′ UTRs, the local regions (101 nt centered on site) containing the edit sites were further enriched for A and U bases (Figure 3.12B, $p = 0.0006$). An example appears in Figure 3.13. These results are consistent with a model in which APOBEC1 targets require a high AU sequence context for efficient editing.

Aside from regional sequence content, other cytidine deaminases in the AID/APOBEC family exhibit strong preferences for particular bases immediately neighboring their editing targets (Beale et al., 2004). Though potential preferences for APOBEC1 have been implied by *in vitro* RNA editing experiments (Backus and Smith, 1992; Chen et al., 1990; Shah et al., 1991), none have been rigorously investigated, likely due to the enzyme's perceived specificity for a single mRNA substrate. In alignments of the 3′ UTR edit sites identified here, almost all of the edited cytidines were immediately flanked by A or U bases at the -1 and +1 position (Figure 3.14, $p = 7 \times 10^{-5}$). There were no significant nucleotide preferences at the -4,-3,-2 and +2,+3,+4 positions relative to the editing site.

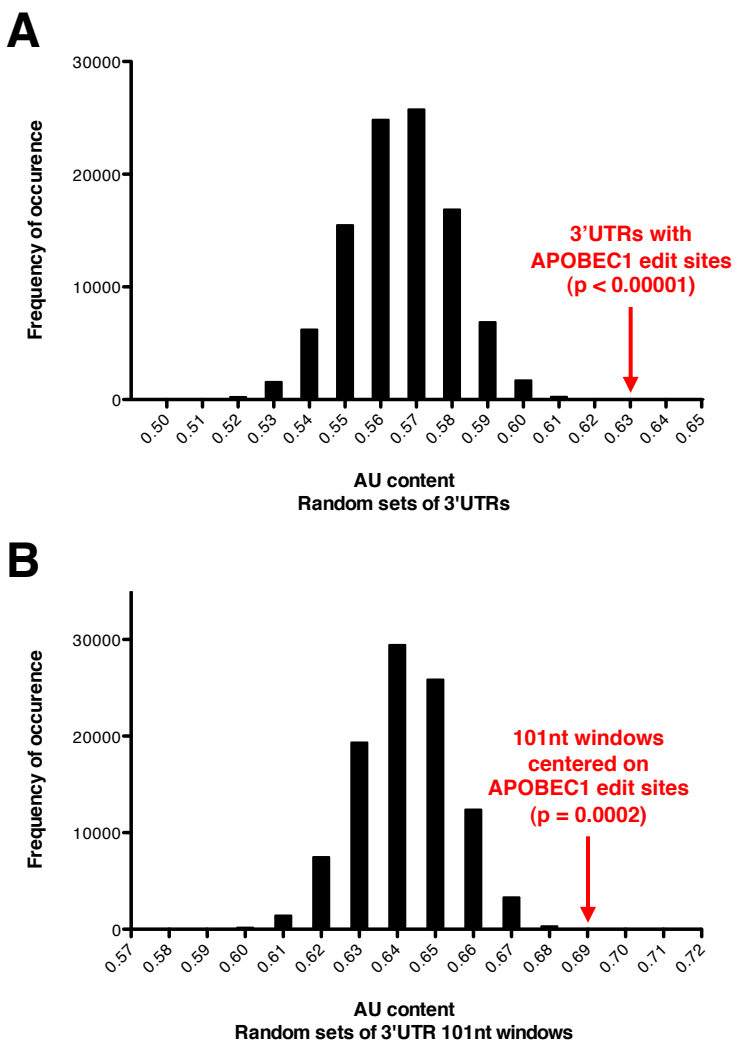


Figure 3.12. Sequence features of APOBEC1 edit sites: AU content. (A) AU content of random sets of 3′ UTRs. The AU content of the set of the 29 edit site-containing 3′ UTRs is 0.63. AU content was computed for each of 100,000 random sets of 29 3′ UTRs, and values were always less than 0.63. Therefore, $p < 0.00001$. (B) AU content of random 101 nt windows within APOBEC1 editing site-containing 3′UTRs. The AU content of the 30 edit site-containing windows is 0.69. The AU content for each of 100,000 random sets of 30 windows was computed and values were greater than or equal to 0.69 in only 0.02% of cases. Therefore, $p = 0.0002$.

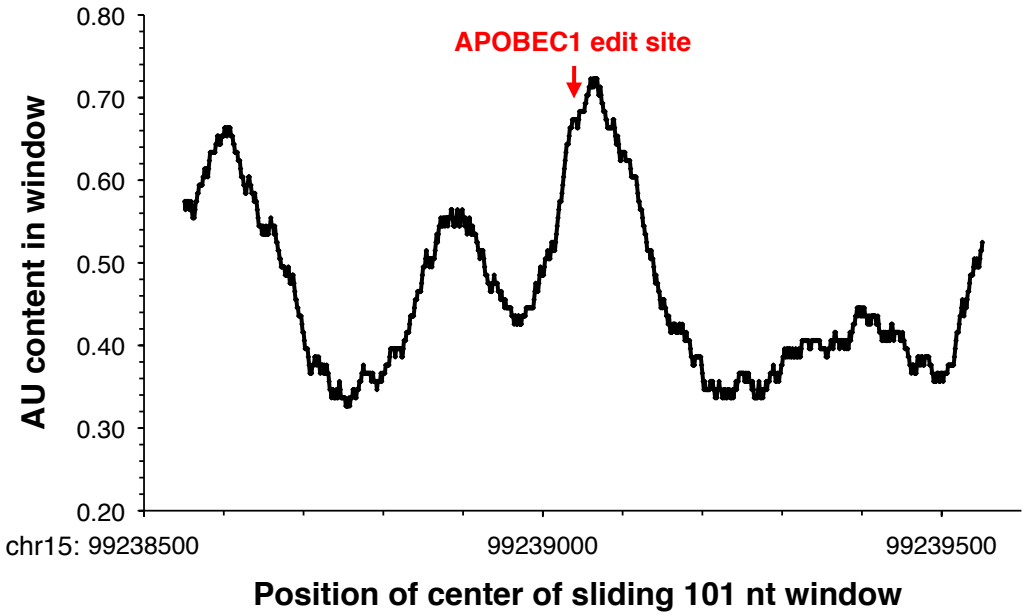
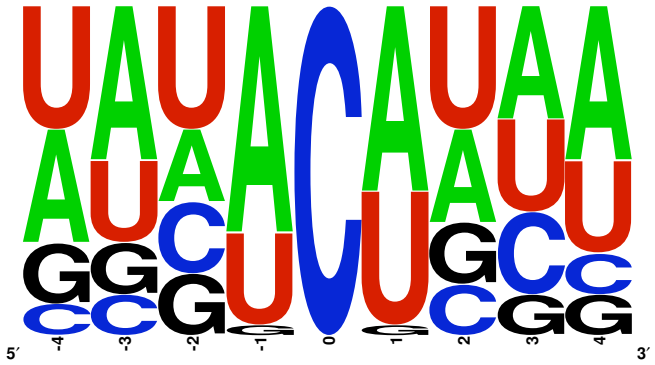


Figure 3.13. AU content in a sliding 101 nt window in the *Tmbim6* 3′ UTR. The AU content peaks near the editing site, chr15:99239051(+).

A



B

5′	-4	-3	-2	-1	Editing Site	+1	+2	+3	+4	3′
A	11	15	7	22	0	18	9	11	15	
C	3	4	7	0	32	0	5	8	4	
G	6	5	6	1	0	1	6	4	4	
T	12	8	12	9	0	13	12	9	9	
P-Value (A or U)	0.4	0.4	0.9	0.00008		0.00008	0.7	0.8	0.3	

Figure 3.14. Sequence features of APOBEC1 edit sites: Flanking nucleotides. (A) Frequency plot of bases flanking the APOBEC1 editing sites, aligned on the target cytidine. (B) Base counts of nucleotides flanking the APOBEC1 editing sites. The nucleotides immediately adjacent to the target cytidine tend overwhelmingly to be A or U. P-values were computed using the binomial test. For a given column, the P-value is the probability that the skew towards A or U is a random occurrence.

DNA and RNA binding proteins often recognize and bind to specific sequence motifs in their molecular targets. To ascertain whether the APOBEC1 editing targets identified here share a common sequence element potentially important for editosome recognition, the Multiple Em for Motif Elicitation (MEME) algorithm (Bailey and Elkan, 1994) was used to analyze the sequence regions (101 nt centered on target C) surrounding the editing sites (Figure 3.15). MEME analysis revealed a significant 10 nt motif (log likelihood ratio = 157, E-value = 8.8×10^{-1}, compared to 65 and 3×10^{2} for shuffled sequence control) in regions adjacent to most (21/31) editing sites. Next, the motif consensus sequence, WRAUYANUAU, was used to manually align the edit site-containing sequences (Figure 3.16). Close or exact consensus motif matches were present downstream (3′) of almost every editing site, with most (24/32) appearing 4-6 nt from the target cytidine. Of note, the consensus motif also matches the first 10 nt of the apoB mooring sequence, which is 5 nt downstream of its editing site (Backus and Smith, 1992; Shah et al., 1991).

Taken together, these results indicate that 3′ UTR targets of APOBEC1 editing are generally in AU-rich regions, immediately flanked by A or U nucleotides, and appear approximately 4-6 nt upstream of a mooring motif. As these features are similar to those in the apoB mRNA (Smith et al., 2005), the 3′ UTR sites may be edited by a similar mechanism.

3.2.5. Sequence features are predictive for APOBEC1 editing in 3′ UTRs

The set of newly identified target sites and their characteristic sequence features described above provide a refined list of criteria for sequences edited by APOBEC1. Based on these findings, an APOBEC1 editing "sequence pattern"

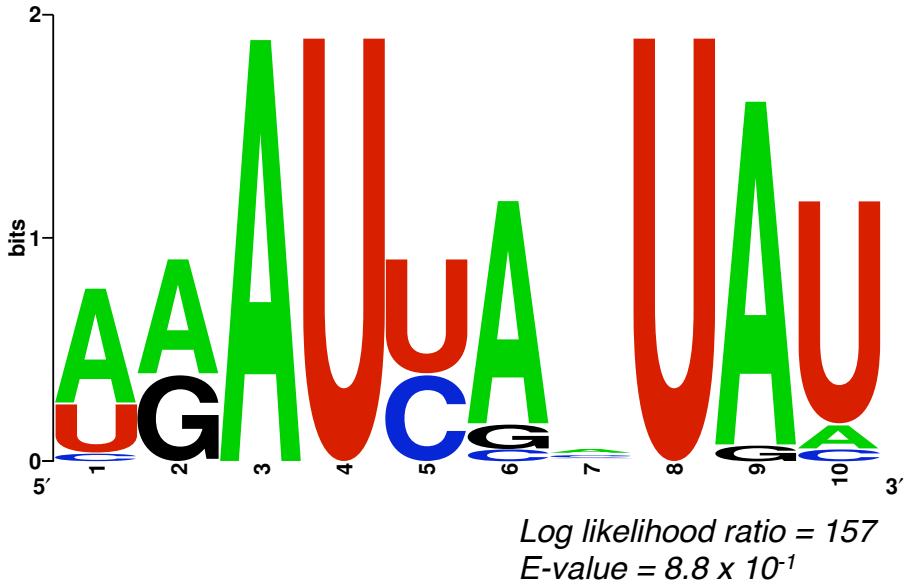


Figure 3.15. Sequence motif identified by MEME analysis of regions flanking APOBEC1 editing sites. This motif was derived from 21 of the 31 sequences analyzed. Log likelihood ratio (157) and E-value (8.8 x10^{-1}) are significant as compared to the "best" motif of a shuffled sequence control (Log likelihood ratio 65, E-value 3 x 10^{2}).

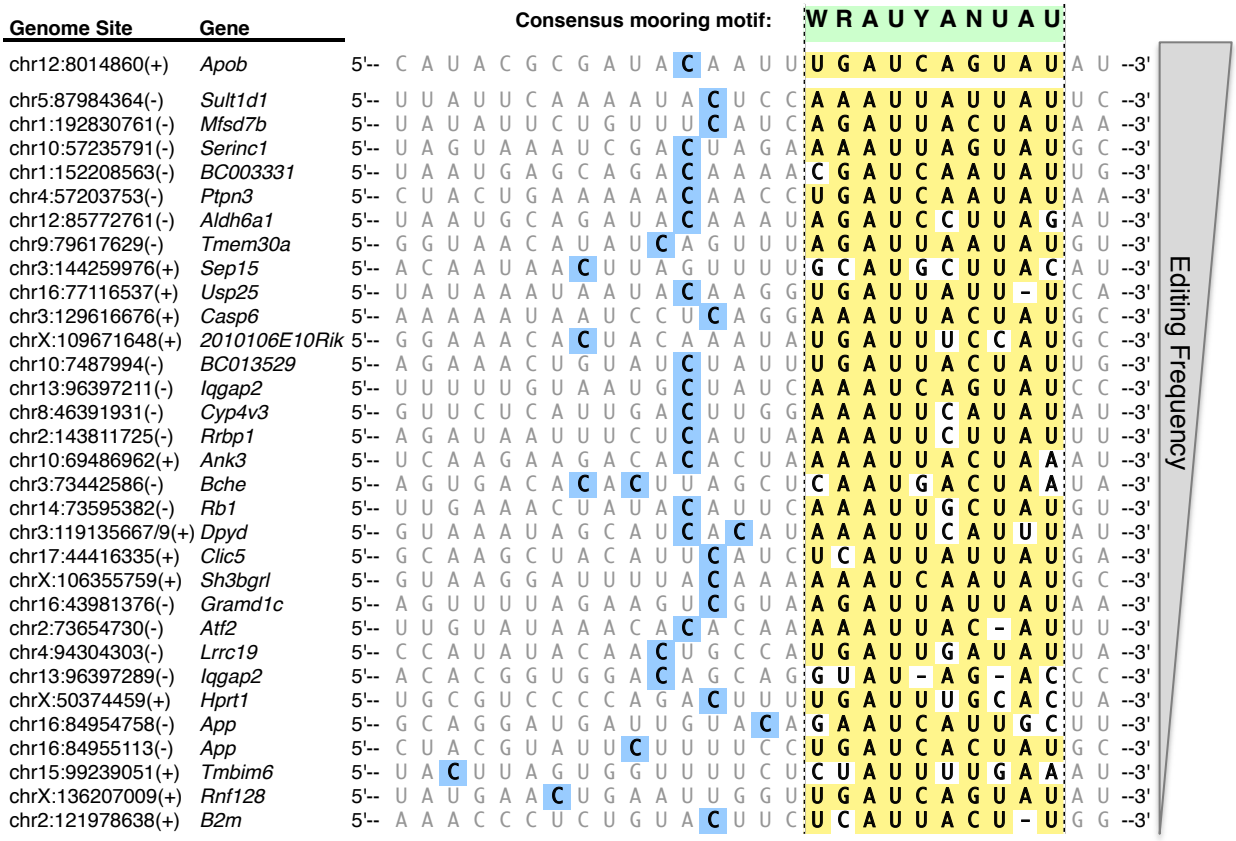


Figure 3.16. Alignment of APOBEC1 target sequences by consensus sequence motif. Edited cytidines are shaded blue. Yellow shading indicates a match to the consensus sequence motif, represented in green. Nearly every editing site is adjacent to at least a partially matched motif, most (24/32) within 4-6 nt.

was composed, consisting of a cytidine flanked on both sides by either A or U and followed by an appropriately spaced mooring motif ($WCWN_{2-4}$WRAUYANUAU, Figure 3.17A). In order to evaluate the distribution of potential APOBEC1 editing targets throughout the transcriptome, all mouse RefSeq exons were searched for this sequence pattern. As enumerated in Figure 3.17B, nearly 400 examples of this pattern were found in mouse mRNAs, 181 of which occurred in transcripts expressed in small intestine enterocytes (RNA-Seq analysis, RPKM ≥ 1.0). Bypassing the comparative editing screen workflow, the sites were directly examined in the wild-type RNA-Seq read sequences at these sites for evidence of RNA editing. Of the 74 patterns located in 3′ UTRs with read coverage (≥3 wild-type reads), detected C/T mismatches indicative of editing were detected at 32 sites. Of the 34 patterns present in coding exons covered by RNA-Seq reads, only the apoB site displayed evidence of editing (Figure 3.17C). A subset of these sites (7 in coding sequences, 14 in 3′ UTR sequences) was additionally examined by standard Sanger sequencing (Figure 3.18). Results confirmed C-to-U editing in 9 of the 3′ UTR sites but none of the coding sequence sites. Though many of the sites described here were not detected in the RNA-Seq screen due to insufficient read coverage in the *apobec1*$^{-/-}$ library and/or relatively low editing frequencies, the APOBEC1 sequence pattern derived from the initially identified targets was predictive for additional APOBEC1 3′ UTR editing sites.

These results suggest that while the APOBEC1 sequence pattern supports editing at numerous sites in transcript 3′ UTRs, it is not targeted when present in coding sequences. The pronounced exception of apoB raises questions about the

A

APOBEC1 editing site pattern

5'-**WCWN$_{2\text{-}4}$WRAUYANUAU**-*3'*

B

Exon type:	5' UTR	CDS	3' UTR
RefSeq transcripts	8	118	263
Expressed in enterocytes (RPKM ≥1.0)	1	45	135
wild-type RNA-Seq read coverage ≥ 3	0	34	74
Sites with C-T mismatch reads (> 0.075)	NA	1	32

C

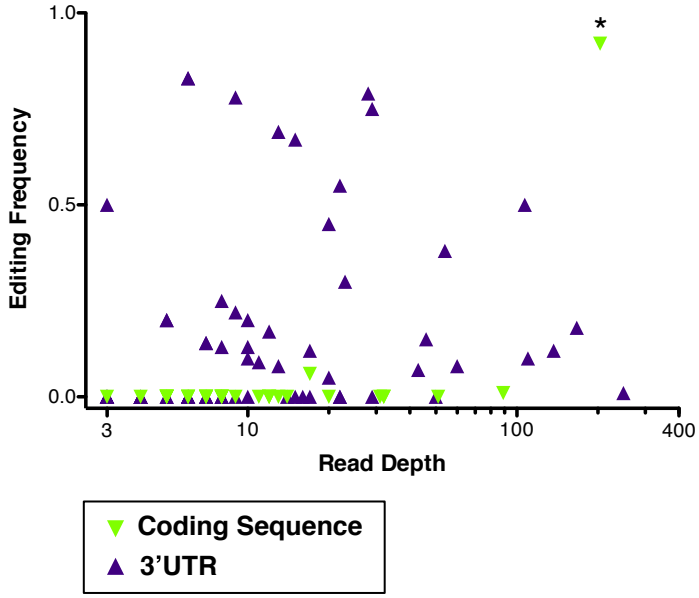


Figure 3.17. Sequence pattern prediction of APOBEC1 mRNA editing sites. (A) APOBEC1 editing site pattern used to search for additional targets in RefSeq transcripts. (B) Occurrences of APOBEC1 editing site pattern in RefSeq transcripts by type, listed by intestinal epithelium expression level and wild-type RNA-Seq read coverage. (C) Editing frequency at predicted APOBEC1 target sites as evaluated by wild-type read content. With the exception of the apoB mRNA, no evidence of editing associated with the APOBEC1 sequence pattern was observed in coding sequences. * *apoB* editing site.

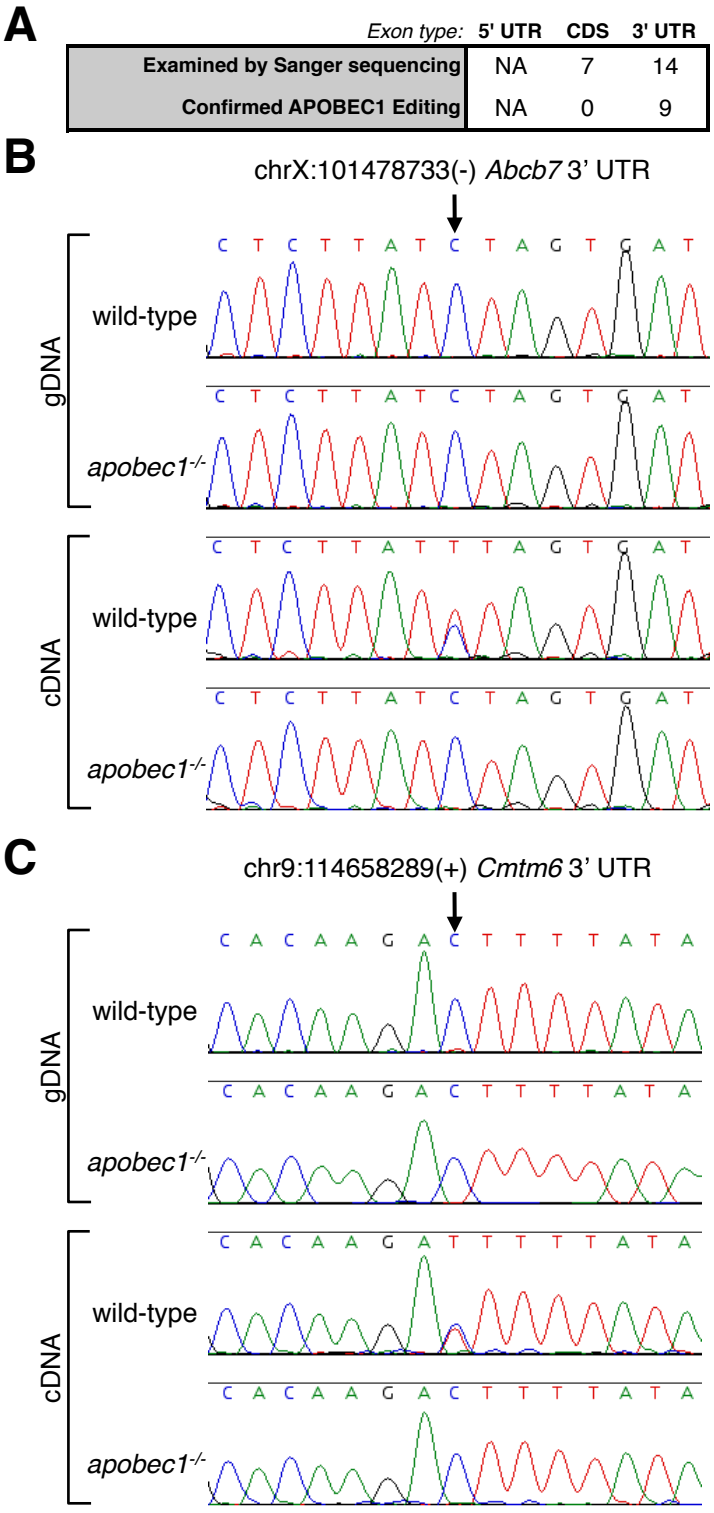


Figure 3.18. Validation of candidate APOBEC1 mRNA editing sites predicted by sequence pattern search. (A) Validation sequencing statistics. Editing was detected in predicted 3' UTRs, but not in coding sequences. (B-C) Representative examples of conventional Sanger sequencing chromatograms for wild-type and *apobec1*$^{-/-}$ genomic DNA and cDNA at predicted editing sites are shown.(A) chrX:101478733(-) in the *Abcb7* transcript, and (B) chr9:114658289(+) in the *Cmtm6* transcript.

mechanism of APOBEC1 sequence recognition and localization as well as the role for editing in 3′ UTRs.

3.2.6. APOBEC1 edit sites within evolutionarily conserved regions

Compared to other non-coding sequences, functional elements within 3′ UTRs are more likely to be conserved throughout evolution. (Duret et al., 1993; Lipman, 1997). Upon inspection of the APOBEC1 editing sites, it appeared that many occur within regions of considerable phylogenetic conservation. Two examples are presented in Figure 3.19A and Figure 3.19B. In order to systematically assess the conservation of sequence regions containing APOBEC1 editing sites, a comparative evolutionary analysis was performed. Conservation scores (phastCons scores for placental mammals (Siepel et al., 2005)) of 101 nt windows centered on the initially identified editing sites were compared to random 101 nt windows within the same 3′ UTRs (Figure 3.19C). As a set, the regions containing APOBEC1 editing sites were significantly more conserved ($P = 0.01$), suggesting that these sequences may be of functional importance.

Furthermore, in multispecies sequence alignments, it appeared that the APOBEC1 editing sites identified in mouse are often fixed as a C base (G base for (-)-strand transcripts) in mammalian evolution (Figure 3.20). In addition, it seemed that when deviating from a C, other genomes most often contain a T (A for (-)-strand transcripts) at the aligned site. This is not unexpected, as transition mutations are significantly more likely than transversion mutations, likely as a consequence of DNA synthesis and proofreading mechanisms (Collins and Jukes, 1994). However, given the C-to-U converting function of APOBEC1, I reasoned that an overrepresentation of C and T bases at genomic sites could

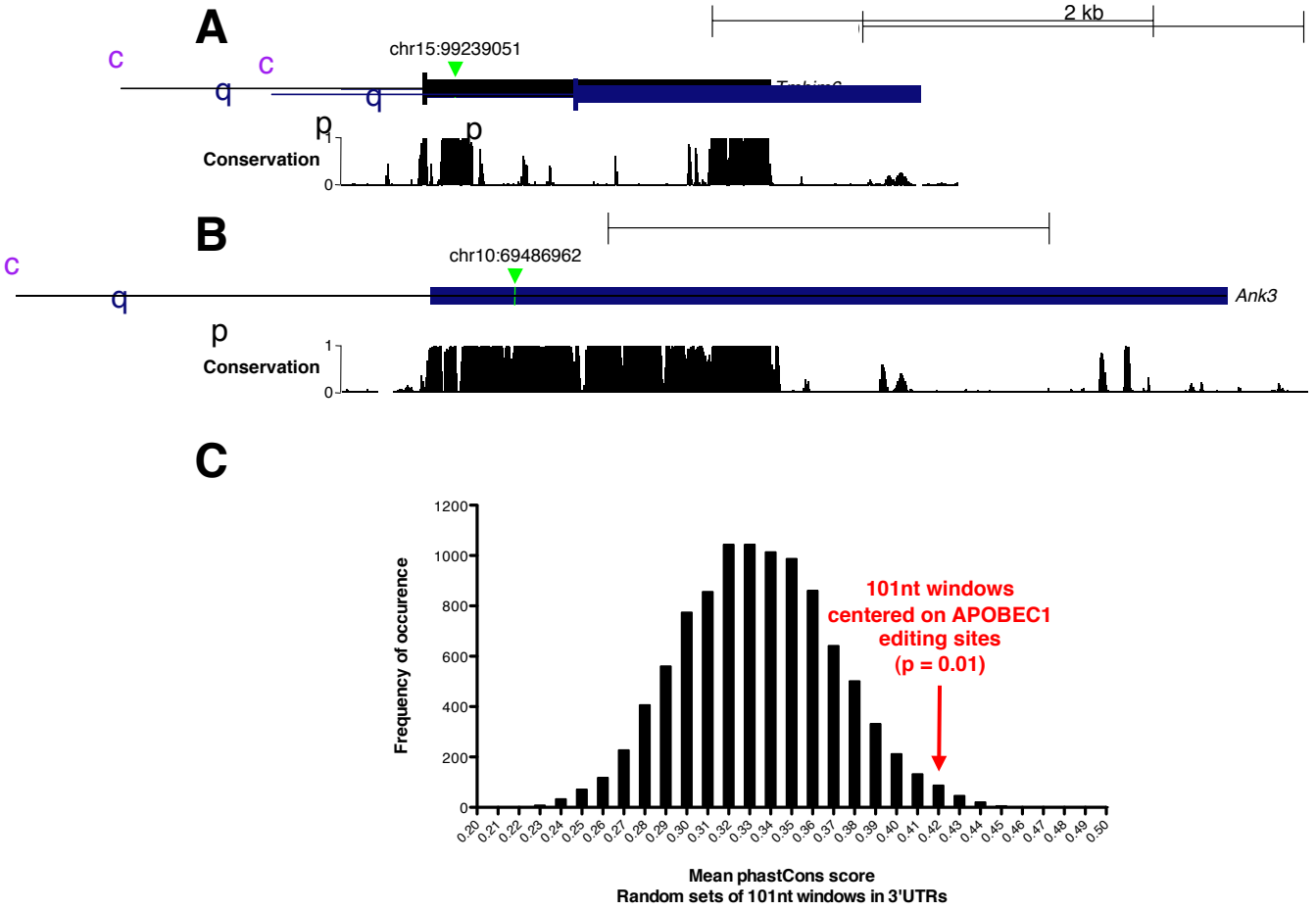


Figure 3.19. Phylogenetic conservation of regions containing APOBEC1 mRNA editing sites. (A and B) Examples of APOBEC1 editing sites within well-conserved regions. Blue bars represent transcript 3′ UTRs. Conservation plots depict phastCons scores for placental mammal multi-alignments. Editing sites are indicated by green arrows. (A) chr15:99239051(+) in the *Tmbim6* transcript, (B) chr10:69486962(+) in the *Ank3* transcript. (C) Phylogenetic conservation for random sets of 101 nt windows within edit site-containing 3′ UTRs, as represented by mean phastCons scores for placental mammal multi-alignments. The mean phastCons score of the 30 edit site containing windows is 0.42. The mean phastCons scores for each of 10,000 random sets of 30 windows was computed and values were greater than or equal to 0.42 in only 1% of the cases. Therefore, p = 0.01.

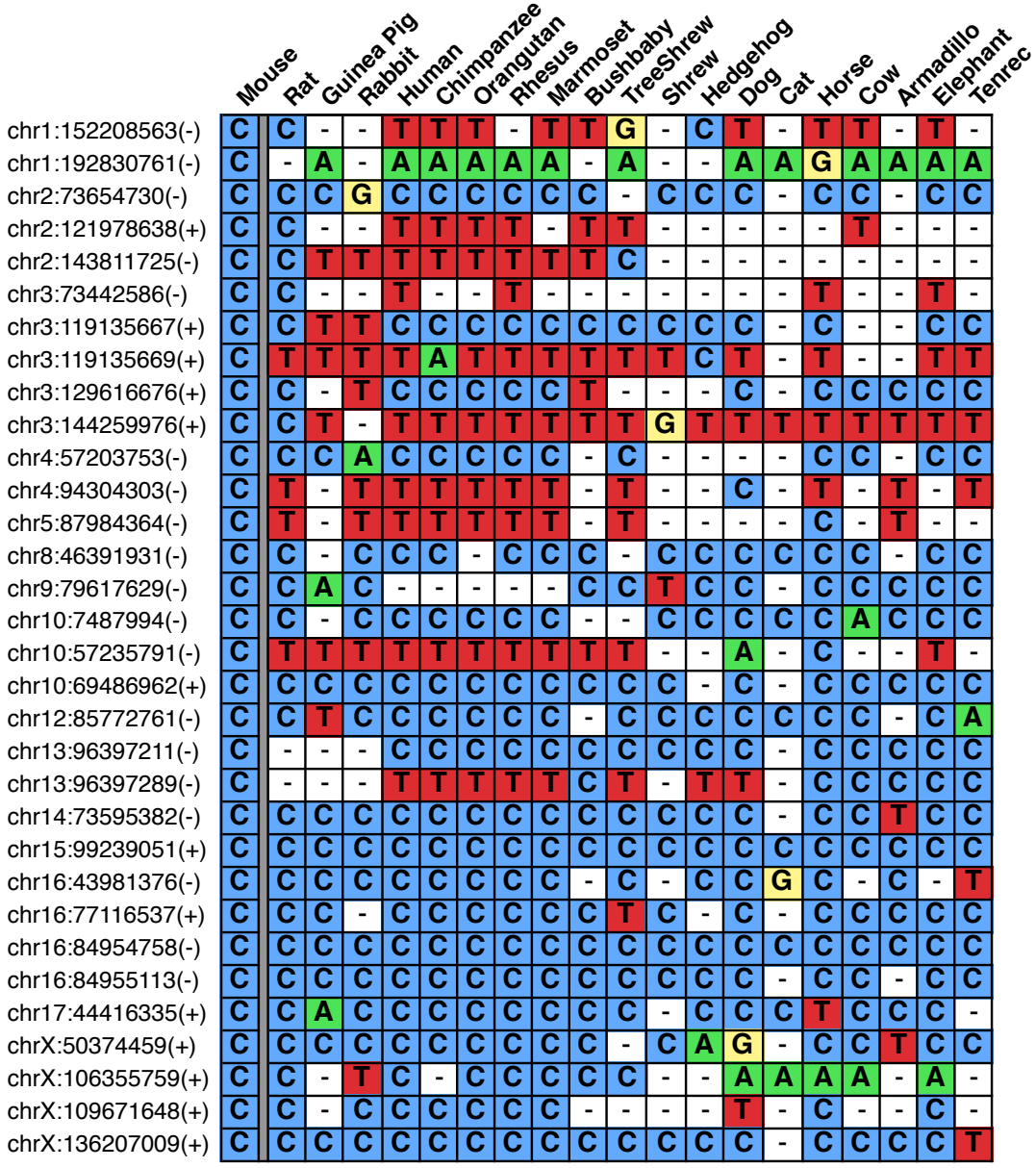


Figure 3.20. Mouse APOBEC1 edit sites in placental mammal genome multi-alignments. Multi-alignments of mouse APOBEC1 editing sites with 19 other placental mammal genomes are shown. When an orthologous nucleotide in another mammal differs from the C in mouse, it is most often a T. This pattern is statistically significant ($p = 0.03$).

reflect functional sequence flexibility recapitulated in some species at the RNA level by APOBEC1. The apparent C-T bias observed in multispecies alignments was quantitatively assessed by comparing the frequency of C-T bases occurring at positions aligned to mouse APOBEC1 editing sites to alignments at comparable random sites. Mammalian genomic alignments to mouse APOBEC1 editing sites are significantly more likely to contain C or T residues ($P = 0.03$), perhaps indicating a functional importance for either base.

In summary, these results indicate that APOBEC1 site-specifically edits many mRNA transcripts other than apoB in small intestinal enterocytes and suggest additional roles for APOBEC1 beyond its function in apolipoprotein regulation.

CHAPTER 4: DISCUSSION

The specific C-to-U base modification of apoB mRNA by APOBEC1 was the first example of mRNA editing observed in mammals (Chen et al., 1987) (Powell et al., 1987). Since this discovery, extensive investigation has provided many details about the cofactors, evolution, regulation, and mechanism of APOBEC1 mRNA editing. However, aside from the NF1 transcript in a subset of tumors (Mukhopadhyay et al., 2002; Skuse et al., 1996), no other physiological targets for APOBEC1 editing have been described. The identification of additional mRNA editing sites has been limited, in part, by the technical challenge of detecting single nucleotide changes within entire transcriptomes.

Recent technological advances now allow for DNA sequencing at a previously unprecedented scale. This thesis presents the development of a novel methodology for the identification of mRNA editing sites by whole transcriptome sequencing. The application of this RNA-Seq screening approach to small intestine enterocytes revealed previously undescribed APOBEC1 editing in numerous mRNA transcripts. With the exception of the apoB mRNA, all newly identified APOBEC1 editing sites are within transcript 3′ UTRs. All of these sites share several characteristic sequence features, including an AU-rich sequence context, a preference for A or U nucleotides immediately adjacent to the edited cytidine, and a downstream APOBEC1 mooring motif similar to that described for the apoB transcript. These features proved to be predictive for APOBEC1 editing, and led to the identification of additional APOBEC1 3′ UTR sites that were not detected by the initial RNA-Seq screen. Taken together, the results presented here dramatically expand the list of validated APOBEC1

mRNA editing targets in small intestinal enterocytes. The recognition of these editing events and their localization to 3′ UTRs raises many questions about functional consequences and potential physiological roles for APOBEC1 beyond lipoprotein regulation.

4.1. APOBEC1 mRNA editing in transcript 3′ UTRs

In the case of apoB, the functional outcome of APOBEC1 editing can be inferred by basic sequence analysis: conversion of a glutamine codon (CAA) to a STOP codon (UAA) within a protein coding sequence results in translation of a truncated apoB-48 product. However, predicting the function of 3′ UTR sequences pre- and/or post- editing is not as straightforward.

Though transcript 3′ UTRs do not encode protein sequence, they have been shown to regulate many aspects of mRNA stability and translational efficiency. 3′ UTRs are the primary site of gene regulation by miRNAs (Chi et al., 2009; Grimson et al., 2007), which can mediate transcript degradation and/or translational suppression. 3′ UTRs can also contain sequence motifs and/or secondary structures important for the recruitment of various regulatory RNA binding proteins. For example, AUUUA pentamers known as AU Rich Elements (AREs) are associated with transcript instability mediated by ARE-binding proteins (Barreau et al., 2005). The 3′ UTRs of transcripts encoding cytokines and other immune mediators often contain variable numbers of AREs, which have been shown to participate in the temporal regulation of inflammatory gene expression (Hao and Baltimore, 2009). Other transcripts can be regulated by structural elements of their 3′ UTRs. For example, the *VEGFA* transcript 3′ UTR contains an RNA secondary structure that undergoes a conformational change in

response to environmental signals and consequently regulates VEGFA protein expression (Ray et al., 2009). Additional regulatory functions attributed to 3′ UTR sequences include mRNA subcellular localization (Jambhekar and Derisi, 2007; Jansen, 2001), translational selenocysteine incorporation (Berry et al., 1991), and transcript retention in the nucleus (Larocque et al., 2002).

The distribution of APOBEC1 sites in coding sequence (apoB) as well as transcript 3′ UTRs is reminiscent of mRNA editing by ADARs. Much like C-to-U deamination in apoB, some initial examples of A-to-I modification in mRNA were observed in tissue-specific transcripts (GluR, in the brain), which have protein coding sequences modified as a consequence of editing (Lomeli et al., 1994). Similar coding changes have been observed in several other neuronal transcripts (reviewed in Bass, 2002). However, recent bioinformatic (Levanon et al., 2004) and ultra high-throughput sequencing analyses (Li et al., 2009b) have demonstrated that most A-to-I RNA editing occurs in non-coding RNA sequences, especially transcript 3′ UTRs. As is the case for APOBEC1, ADAR editing varies in efficiency for different target transcripts (Li et al., 2009b). Though the functional consequences for most of these editing events remain largely unknown, several targets have been examined in some detail. Evidence suggests that A-to-I editing in 3′ UTR sequence can induce nuclear retention of transcripts (Chen et al., 2008), target mRNA cleavage (Osenberg et al., 2009), and potentially modify miRNA target sites to modulate gene expression (Borchert et al., 2009; Liang and Landweber, 2007).

The recent evidence that ADAR editing of transcript 3′ UTRs can affect post-transcriptional regulation of RNA provides illustrative examples of how changes in non-coding sequence can impact genetic output. Might APOBEC1

editing of 3′ UTRs have similar functional consequences? Based on the sequence context of the identified edit sites, a number of regulatory possibilities can be envisioned. First, APOBEC1 editing events in four transcript 3′ UTRs are predicted to generate new AREs. For example, in the *Tmbim6* transcript 3′ UTR, APOBEC1 editing converts ACUUA to AUUUA, a canonical ARE. As the number of AREs contained within a transcript 3′ UTR has been shown to inversely correlate with mRNA stability (Hao and Baltimore, 2009), generating additional AREs could destabilize an edited transcript. Furthermore, because the edited 3′ UTRs are high in overall AU content, several already contain AREs other than those generated by APOBEC1 editing. It is possible that introduction of an additional ARE could contribute to the fine-tuning of transcript stability. Of note, though editing was not observed, APOBEC1 has been reported to bind and stabilize AU-rich 3′ UTRs (Anant and Davidson, 2000; Anant et al., 2004). Perhaps APOBEC1 can modulate mRNA stability through editing-dependent and –independent mechanisms.

Alteration of miRNA targets represents another possible functional consequence of 3′ UTR editing. 3′ UTRs represent the principle targets of transcript regulation by miRNAs. More than 35% of APOBEC1 editing sites are located within sequences that match the seed targets of known miRNAs. Cytidine deamination at these sites modifies target sequences and could potentially abolish miRNA binding. Conversely, APOBEC1 editing could introduce new miRNA seed target sequences, or shift existing targets to sequences that recruit different miRNAs. It should be noted that miRNA targeting is enhanced within regions rich in A and U nucleotides (Grimson et al., 2007), a prominent feature of APOBEC1 editing sites. Though most edited

transcripts were measured at similar levels by RNA-Seq profiling, several were 1.5 – 2.0 fold differentially expressed (wild-type vs. *apobec1*$^{-/-}$), differences consistent with miRNA regulation. Furthermore, as miRNA targeting in vertebrates may primarily affect translation rather than mRNA degradation (Jackson and Standart, 2007; Standart and Jackson, 2007), alterations due to editing might not be apparent at the transcript level.

Without flexible mouse enterocyte models that can manipulated *in vitro*, direct experimental evidence for the functional and physiological relevance of these editing events would require detection of altered translational outcomes in APOBEC1-expressing enterocytes *in vivo*. Furthermore, *apobec1*$^{-/-}$ epithelial cells accumulate triacylglycerol lipids due to apoB-related deficiencies in chylomicron formation (Kendrick et al., 2001). Therefore, direct regulatory effects due to the absence of 3′ UTR editing of various target transcripts are difficult to evaluate, as they may be obscured by the indirect cellular effects of the absence of apoB editing on lipid metabolism. For these technical reasons, experimental evidence for the importance of APOBEC1 editing has been elusive. However, the localization of many APOBEC1 edit sites within regions conserved throughout mammalian evolution implies potential functional relevance. Indeed, the APOBEC1 editing sites are significantly more likely to occur within highly conserved sequence regions than in other, less conserved 3′ UTR sequences. Furthermore, though analysis was limited by a relatively small sample set (n = 32 sites), editing sites in mouse transcripts tend to be maintained as genomic Cs or Ts throughout mammalian evolution significantly more so than would be expected at background phylogenetic transition rates (p=0.03). This observation could be indicative of a mechanism whereby APOBEC1 "corrects" a genomic

cytidine mutation to an appropriate T within specific tissues and/or under certain conditions. Alternatively, APOBEC1 editing may provide a means of "genetic dosing" at specific nucleotide positions, thereby adding adaptive flexibility at the transcript level to sequences hard coded in the genome, as has been recently proposed for ADARs (Gommans et al., 2009).

4.2. Sequence features of APOBEC1 mRNA editing targets in 3′ UTRs

Without evidence of additional mRNA targets, most characterizations of the sequence and structural requirements for APOBEC1 editing have been based on the apoB transcript. Early studies demonstrated that editing required a downstream mooring sequence separated from the target cytidine by a short spacer element (Backus and Smith, 1992; Shah et al., 1991). The size of the spacer element was found to be somewhat flexible; acceptable lengths ranged from 4 to 7 nt, with an optimal distance of 5 nt. Similarly, editing activity can tolerate mooring sequence point mutations at some positions but not others (Shah et al., 1991). However, these analyses were limited only to transversion mutations, and may have overlooked additional sequence flexibility.

Analysis of the numerous mRNA targets identified here demonstrate flexibility in the sequence requirements for APOBEC1 editing consistent with, but not apparent in, previous studies of apoB. The consensus mooring motif derived from the newly identified targets implies some rigid sequence constraints, notably at position 3 (A) and position 4 (U). These apparent constrains are consistent with the *in vitro* analysis of the apoB mRNA; point mutation at these positions completely abrogated editing by APOBEC1 (Shah et al., 1991). However, it appears that other positions tolerate different nucleotides

than those described for apoB. For example, most targets contain a G or an A at position 2, suggesting a purine requirement. Similarly, most targets contain a C or a U at position 5, indicating a likely pyrimidine constraint. The flexibility limited to nucleotides with similar aromatic rings may reflect certain structural requirements of the APOBEC1 active site and/or the RNA binding components of the editosome. In current mechanistic models for apoB mRNA editing, ACF binds the mooring sequence, and its appropriate spacing ensures proper placement of the target cytidine in the APOBEC1 active site (Maris and Allain, 2009). The recognition of similar mooring sequences in 3' UTR targets suggests that editing of these sites proceeds by a similar mechanism. The strong preference observed for A and U nucleotides immediately flanking the edit sites has not been previously described, though the edited cytidine in apoB mRNA is bordered by As. Perhaps these flanking nucleotides fulfill an additional structural or mechanistic requirement for proper positioning and deamination in the APOBEC1 active site.

Feature characterization of multiple APOBEC1 mRNA edit sites has also provided a flexible sequence pattern for robust bioinformatic prediction of additional transcript targets. Previous computational attempts to identify APOBEC1 edit sites in sequence databases were limited by patterns derived solely from the apoB mRNA and related experimentation. For example, sequence queries using weighted matrix motif models developed from apoB mooring sequence point mutagenesis studies identified numerous candidate APOBEC1 editing sites throughout the transcriptome (Smith et al., 2005). However, editing was not detected at any of the sites examined. Therefore, though required for editing, the mooring sequence alone is not adequately

predictive for APOBEC1 mRNA substrates. In contrast, the refined sequence pattern derived from the collection of targets reported here was sufficient to identify numerous additional APOBEC1 sites that were not detected by the RNA-Seq screen. These results further support a model for APOBEC1 target recognition compatible with but more complex than that presumed from apoB mRNA editing.

4.3. APOBEC1 mRNA editing appears to be constrained to 3′ UTRs

The localization of all newly identified editing sites to transcript 3′ UTRs raises questions about the mechanism of apoB mRNA coding sequence editing by APOBEC1. Despite the presence of sequence motifs consistent with APOBEC1 editing within coding and untranslated sequences throughout the transcriptome, RNA-Seq data suggest that APOBEC1 only acts on those targets located in 3′ UTRs. Thus, with regard to APOBEC1 targeting, apoB coding sequence editing appears to be the exception rather than the rule. How is the APOBEC1 editosome targeted to its apparent recognition motif in apoB mRNA and transcript 3′ UTRs but not to similar motifs in other coding sequences? Prior to the identification of the additional editing sites described here, it was proposed that the RNA splicing machinery physically obscures the sizeable editosome from access to most coding exons, restricting editing to a post-splicing, pre-nuclear-export temporal window (Sowden et al., 1996b; Sowden and Smith, 2001). In this model, the apoB target remains accessible because the APOBEC1 target site is located near the midpoint of a particularly large (>7 kb) exon and therefore sufficiently distant from exon-intron boundaries and the spliceosome protein complexes associated with them. Such a mechanism might

allow for the editing of 3′ UTR sites observed here. A similar possibility might involve a more specific exclusion of APOBEC1 from all coding sequences in order to protect them from off-target editing. This exclusionary mechanism could be bypassed at the apoB edit site by an additional targeting factor within the editosome, on the apoB mRNA, or both. It is interesting to note that upon apoB mRNA editing by APOBEC1 the downstream coding sequence becomes a 3′ UTR. This may represent an important link regarding the relationship of this well-known mRNA editing target to the 3′ UTR editing sites described here.

4.4. Functions for APOBEC1 beyond the small intestine

In many mammals, APOBEC1 expression is not restricted to the small intestine. For example, murine APOBEC1 is also expressed in liver, kidney, muscle and spleen (Nakamuta et al., 1995). APOBEC1 expression has been detected in immune cell subsets, including macrophages, dendritic cells and B cells. Furthermore, expression in these cell types is upregulated by TLR stimuli such as LPS and poly(I:C). However, the function of APOBEC1 in these cells is unclear, as the apoB transcript is not expressed. Given the expression pattern in immune cells and the host defense functions of many related AID/APOBEC cytidine deaminases, perhaps APOBEC1 editing functions in the immune system. If true, one might expect a readily observable immune phenotype in *apobec1*$^{-/-}$ mice, which has not been described. However, despite extensive experimentation, *apobec1*$^{-/-}$ animals have been used primarily for the study of apoB mRNA editing and its physiological effects on lipid regulation. Furthermore, an immune phenotype might only be apparent in a specific context, i.e. infection with a particular pathogen. For example, though APOBEC3 is an

effective inhibitor of retroviral infection, *apobec3*$^{-/-}$ mice are healthy and phenotypically normal on gross examination (Mikl et al., 2005). However, upon infection with Moloney murine leukemia virus (Low et al., 2009) or murine mammary tumor virus (Okeoma et al., 2009), *apobec3*$^{-/-}$ mice exhibit deficiencies in viral control as compared to wild-type animals.

The recognition that APOBEC1 editing is not restricted solely to the apoB mRNA raises the possibility that APOBEC1 might edit other transcripts in immune cells. Application of the comparative RNA-Seq screen to the immune cells of steady state and/or infected wild-type and *apobec1*$^{-/-}$ mice will likely provide insight as to potential mRNA editing targets and functions for APOBEC1 in host defense.

4.5. Comparative RNA-Seq screen for the study of mRNA editing: Advantages and disadvantages

The recent advances in sequencing technologies offer new and powerful tools with which to study cellular transcriptomes. Several ultra-high throughput sequencing methods have been applied to RNA editing problems, and each has a distinct set of advantages and drawbacks. Though very effective in identifying mRNA targets of APOBEC1 editing, the comparative RNA-Seq screening approach presented here is no exception. Perhaps most importantly, the comparative screen is not biased by any presuppositions or assumptions regarding potential editing targets; sites are identified strictly through the detection of single nucleotide mismatches. This contrasts with recent ADAR editing studies that used ultra-high throughput sequencing to confirm editing sites predicted by sequence context and/or EST database analysis (Li et al.,

2009b). Furthermore, compared to sequence capture techniques, RNA-Seq is considerably more cost effective in sample preparation and analysis. Next, as a comparative technique with a controlled variable (plus/minus editing activity), editing events can be assigned to a specific enzyme, in this case APOBEC1. Finally, in addition to the nucleotide mismatches indicative of editing, RNA-Seq datasets can provide extensive supplemental information regarding transcript expression, splicing, relative isoform abundance, and promoter usage. Though not required for edit site identification, such information can be useful in interpreting the functional consequences of mRNA modifications.

The primary disadvantage of an RNA-Seq approach to editing is that transcripts expressed at low levels are likely to be underrepresented in sequencing datasets. Sequencing coverage is directly proportional to transcript expression level (Mortazavi et al., 2008), and transcripts with insufficient coverage cannot be interrogated for potential mismatches. This problem is evident in the identification and validation of several sequence-predicted editing targets that were overlooked by the RNA-Seq screen due to inadequate read coverage. This limitation can be overcome by sequencing RNA-Seq libraries at additional read depth. Next, as a comparative screen, this technique requires "editing enzyme-competent" and "editing enzyme-deficient" samples. As demonstrated for APOBEC1, congenic knockout mice are an ideal source of experimental material. However, as is the case for ADAR1 (Hartner et al., 2004; Wang et al., 2004b), certain RNA editing enzymes may be essential for development and viability. Furthermore, a requirement for genetically modified organisms eliminates the possibility of studies with human tissue. These issues could be addressed by conditional deletion and/or RNAi knockdown strategies.

Finally, in its current form, the comparative screen uses only polyA$^+$ mRNA in RNA-Seq library preparation. As a result, it cannot detect editing events in transcripts that are not polyadenylated, including various non-coding RNAs and miRNAs. This limitation can be circumvented through different RNA isolation protocols, such as negative selection methods that deplete ribosomal RNA from total RNA preparations.

4.6. Comparative RNA-Seq mRNA editing screen: Additional applications

The successful development and application of an RNA-Seq screening approach to identify mRNA editing sites provides an opportunity to address many additional questions about RNA editing enzymes and their targets. This methodology can be easily adapted and applied to a variety of editing enzymes in diverse organisms and tissues. As described above, a characterization of C-to-U editing in immune cells would be helpful in understanding possible functions for APOBEC1 in host defense. The screen could also be applied to cytidine deaminases without known targets; a whole-transcriptome comparison of wild-type and *apobec2*$^{-/-}$ muscle (Mikl et al., 2005) might elucidate the long sought-after substrate(s) for APOBEC2. An investigation of potential AID mRNA editing is another intriguing possibility. Upon its initial discovery, AID was thought to be an RNA editor on account of its homology to APOBEC1 (Muramatsu et al., 1999). Though since demonstrated to act on genomic DNA, the possibility of additional RNA substrates for AID has not been ruled out (Shivarov et al., 2008). However, though a complete analysis is pending, a preliminary RNA-Seq screening experiment using *aicda*$^{-/-}$ B cells suggests that AID does not edit mRNA during immunoglobulin CSR.

Though effective for APOBEC1, the comparative RNA-Seq screen is not limited to editing targets of cytidine deaminases. Identifying RNA editing events in the recently developed conditional ADAR1 deletion mouse system (XuFeng et al., 2009) could supplement and assign specificity to many recently described A-to-I editing targets. In addition, there are many outstanding questions regarding RNA editing in other non-mammalian species that could be similarly addressed. For example, C-to-U mRNA editing was recently observed in *C. elegans*; the extent of its activity and function remain largely unknown (Wang et al., 2004a). Finally, this methodology is not strictly limited to the study of RNA editing. A similar transcriptomics workflow can be utilized for any comparative evaluation of single nucleotide differences in mRNA. For example, "expressed mutations" (i.e. genomic mutations in transcribed exons) could be evaluated in tumor versus healthy control tissue from the same patient. Such an analysis might reveal point mutations, insertions and/or deletions associated with oncogenesis while simultaneously providing gene expression profiling data. These and other applications illustrate the potential versatility of this transcriptomics methodology.

4.7. Closing remarks

Recent technological advances allow for the study of genomes and transcriptomes on a formerly unprecedented scale. As more data become available, it is increasingly apparent that the transcriptome is far more dynamic and complex than previously anticipated. RNA transcripts can undergo a wide variety of alterations, including splicing, cleavage, base modification and editing. Such mechanisms introduce considerable diversity to the transcriptome and, by

extension, to the expressed proteome. RNA editing is one mechanism by which the expressed information content of a gene can be altered without modifying the genome itself.

As the information encoded in an mRNA transcript can affect a protein's sequence and the regulation of its expression, mRNA editing can impact several different aspects of the cellular proteome. In the case of ADARs, editing in coding sequences gives rise to functionally distinct proteins and editing in 3' UTRs affects transcript regulation and expression. Though long thought monospecific for the apoB coding sequence, the findings presented here demonstrate that APOBEC1 edits numerous mRNAs in small intestinal enterocytes. The localization of these newly identified editing events to 3' UTRs suggests that they may play a role in transcript regulation. These results also imply additional functions for this cytidine deaminase beyond its characterized role in lipid transport, both in small intestinal enterocytes as well as other cell types. Furthermore, the number and diversity of APOBEC1 targets identified provides an example of the additional complexity introduced to a cellular transcriptome by RNA editing. Such informational diversity, whether mediated by APOBEC1 or other RNA editing enzymes, demonstrates one aspect of the flexible and dynamic nature of mRNA transcripts, which were originally thought to be static facsimiles of genetic content. The trancriptomics methodology presented here will be useful in understanding the informational diversity generated by RNA editing and its functional impact in diverse biological systems.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., *et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Anant, S., and Davidson, N.O. (2000). An AU-rich sequence element (UUUN[A/U]U) downstream of the edited C in apolipoprotein B mRNA is a high-affinity binding site for Apobec-1: binding of Apobec-1 to this motif in the 3' untranslated region of c-myc increases mRNA stability. Mol Cell Biol *20*, 1982-1992.

Anant, S., Henderson, J.O., Mukhopadhyay, D., Navaratnam, N., Kennedy, S., Min, J., and Davidson, N.O. (2001a). Novel role for RNA-binding protein CUGBP2 in mammalian RNA editing. CUGBP2 modulates C to U editing of apolipoprotein B mRNA by interacting with apobec-1 and ACF, the apobec-1 complementation factor. J Biol Chem *276*, 47338-47351.

Anant, S., MacGinnitie, A.J., and Davidson, N.O. (1995). apobec-1, the catalytic subunit of the mammalian apolipoprotein B mRNA editing enzyme, is a novel RNA-binding protein. J Biol Chem *270*, 14762-14767.

Anant, S., Mukhopadhyay, D., Sankaranand, V., Kennedy, S., Henderson, J.O., and Davidson, N.O. (2001b). ARCD-1, an apobec-1-related cytidine deaminase, exerts a dominant negative effect on C to U RNA editing. Am J Physiol, Cell Physiol *281*, C1904-1916.

Anant, S., Murmu, N., Houchen, C.W., Mukhopadhyay, D., Riehl, T.E., Young, S.G., Morrison, A.R., Stenson, W.F., and Davidson, N.O. (2004). Apobec-1 protects intestine from radiation injury through posttranscriptional regulation of cyclooxygenase-2 expression. Gastroenterology *127*, 1139-1149.

Aphasizhev, R., Aphasizheva, I., Nelson, R.E., Gao, G., Simpson, A.M., Kang, X., Falick, A.M., Sbicego, S., and Simpson, L. (2003). Isolation of a U-insertion/deletion editing complex from Leishmania tarentolae mitochondria. EMBO J *22*, 913-924.

Athanasiadis, A., Rich, A., and Maas, S. (2004). Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol *2*, e391.

Backus, J.W., and Smith, H.C. (1991). Apolipoprotein B mRNA sequences 3' of the editing site are necessary and sufficient for editing and editosome assembly. Nucleic Acids Res *19*, 6781-6786.

Backus, J.W., and Smith, H.C. (1992). Three distinct RNA sequence elements are required for efficient apolipoprotein B (apoB) RNA editing in vitro. Nucleic Acids Res *20*, 6007-6014.

Baczko, K., Lampe, J., Liebert, U.G., Brinckmann, U., ter Meulen, V., Pardowitz, I., Budka, H., Cosby, S.L., Isserte, S., and Rima, B.K. (1993). Clonal expansion of hypermutated measles virus in a SSPE brain. Virology *197*, 188-195.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol *2*, 28-36.

Barreau, C., Paillard, L., and Osborne, H.B. (2005). AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Res *33*, 7138-7150.

Basilio, C., Wahba, A.J., Lengyel, P., Speyer, J.F., And Ochoa, S. (1962). Synthetic polynucleotides and the amino acid code. V. Proc Natl Acad Sci USA *48*, 613-616.

Bass, B.L. (2002). RNA editing by adenosine deaminases that act on RNA. Annu Rev Biochem *71*, 817-846.

Bass, B.L., and Weintraub, H. (1987). A developmentally regulated activity that unwinds RNA duplexes. Cell *48*, 607-613.

Bass, B.L., and Weintraub, H. (1988). An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell *55*, 1089-1098.

Beale, R.C.L., Petersen-Mahrt, S.K., Watt, I.N., Harris, R.S., Rada, C., and Neuberger, M.S. (2004). Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra in vivo. J Mol Biol *337*, 585-596.

Benne, R., Van den Burg, J., Brakenhoff, J.P., Sloof, P., Van Boom, J.H., and Tromp, M.C. (1986). Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell *46*, 819-826.

Bennett, R.P., Diner, E., Sowden, M.P., Lees, J.A., Wedekind, J.E., and Smith, H.C. (2006). APOBEC-1 and AID are nucleo-cytoplasmic trafficking proteins but APOBEC3G cannot traffic. Biochem Biophys Res Commun *350*, 214-219.

Bernard, A., and Khrestchatisky, M. (1994). Assessing the extent of RNA editing in the TMII regions of GluR5 and GluR6 kainate receptors during rat brain development. J Neurochem *62*, 2057-2060.

Berry, M.J., Banu, L., Chen, Y.Y., Mandel, S.J., Kieffer, J.D., Harney, J.W., and Larsen, P.R. (1991). Recognition of UGA as a selenocysteine codon in type I deiodinase requires sequences in the 3' untranslated region. Nature *353*, 273-276.

Betts, L., Xiang, S., Short, S.A., Wolfenden, R., and Carter, C.W. (1994). Cytidine deaminase. The 2.3 A crystal structure of an enzyme: transition-state analog complex. J Mol Biol *235*, 635-656.

Bishop, K.N., Holmes, R.K., Sheehy, A.M., Davidson, N.O., Cho, S.-J., and Malim, M.H. (2004). Cytidine deamination of retroviral DNA by diverse APOBEC proteins. Curr Biol *14*, 1392-1396.

Blanc, V., Henderson, J.O., Kennedy, S., and Davidson, N.O. (2001a). Mutagenesis of apobec-1 complementation factor reveals distinct domains that modulate RNA binding, protein-protein interaction with apobec-1, and complementation of C to U RNA-editing activity. J Biol Chem *276*, 46386-46393.

Blanc, V., Henderson, J.O., Newberry, E.P., Kennedy, S., Luo, J., and Davidson, N.O. (2005). Targeted deletion of the murine apobec-1 complementation factor (acf) gene results in embryonic lethality. Mol Cell Biol *25*, 7260-7269.

Blanc, V., Kennedy, S., and Davidson, N.O. (2003). A novel nuclear localization signal in the auxiliary domain of apobec-1 complementation factor regulates nucleocytoplasmic import and shuttling. J Biol Chem *278*, 41198-41204.

Blanc, V., Navaratnam, N., Henderson, J.O., Anant, S., Kennedy, S., Jarmuz, A., Scott, J., and Davidson, N.O. (2001b). Identification of GRY-RBP as an apolipoprotein B RNA-binding protein that interacts with both apobec-1 and apobec-1 complementation factor to modulate C to U editing. J Biol Chem *276*, 10272-10283.

Blow, M., Futreal, P.A., Wooster, R., and Stratton, M.R. (2004). A survey of RNA editing in human brain. Genome Res *14*, 2379-2387.

Blow, M.J., Grocock, R.J., van Dongen, S., Enright, A.J., Dicks, E., Futreal, P.A., Wooster, R., and Stratton, M.R. (2006). RNA editing of human microRNAs. Genome Biol *7*, R27.

Blum, B., Bakalara, N., and Simpson, L. (1990). A model for RNA editing in kinetoplastid mitochondria: "guide" RNA molecules transcribed from maxicircle DNA provide the edited information. Cell *60*, 189-198.

Bogerd, H.P., Doehle, B.P., Wiegand, H.L., and Cullen, B.R. (2004). A single amino acid difference in the host APOBEC3G protein controls the primate species specificity of HIV type 1 virion infectivity factor. Proc Natl Acad Sci USA *101*, 3770-3774.

Borchert, G.M., Gilmore, B.L., Spengler, R.M., Xing, Y., Lanier, W., Bhattacharya, D., and Davidson, B.L. (2009). Adenosine deamination in human transcripts generates novel microRNA binding sites. Hum Mol Genet *18*, 4801-4807.

Brar, S.S., Watson, M., and Diaz, M. (2004). Activation-induced cytosine deaminase (AID) is actively exported out of the nucleus but retained by the induction of DNA breaks. J Biol Chem *279*, 26395-26401.

Burns, C.M., Chu, H., Rueter, S.M., Hutchinson, L.K., Canton, H., Sanders-Bush, E., and Emeson, R.B. (1997). Regulation of serotonin-2C receptor G-protein coupling by RNA editing. Nature *387*, 303-308.

Casey, J.L., and Gerin, J.L. (1995). Hepatitis D virus RNA editing: specific modification of adenosine in the antigenomic RNA. J Virol *69*, 7593-7600.

Cattaneo, R. (1994). Biased (A-->I) hypermutation of animal RNA virus genomes. Curr Opin Genet Dev *4*, 895-900.

Cattaneo, R., Schmid, A., Eschle, D., Baczko, K., ter Meulen, V., and Billeter, M.A. (1988). Biased hypermutation and other genetic changes in defective measles viruses in human brain infections. Cell *55*, 255-265.

Cavalier-Smith, T. (1997). Cell and genome coevolution: facultative anaerobiosis, glycosomes and kinetoplastan RNA editing. Trends Genet *13*, 6-9.

Chan, L. (1992). Apolipoprotein B, the major protein component of triglyceride-rich and low density lipoproteins. J Biol Chem *267*, 25621-25624.

Chang, F.L., Chen, P.J., Tu, S.J., Wang, C.J., and Chen, D.S. (1991). The large form of hepatitis delta antigen is crucial for assembly of hepatitis delta virus. Proc Natl Acad Sci USA *88*, 8490-8494.

Chaudhuri, J., Tian, M., Khuong, C., Chua, K., Pinaud, E., and Alt, F.W. (2003). Transcription-targeted DNA deamination by the AID antibody diversification enzyme. Nature *422*, 726-730.

Chen, C.X., Cho, D.S., Wang, Q., Lai, F., Carter, K.C., and Nishikura, K. (2000). A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. RNA *6*, 755-767.

Chen, H., Lilley, C.E., Yu, Q., Lee, D.V., Chou, J., Narvaiza, I., Landau, N.R., and Weitzman, M.D. (2006). APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. Curr Biol *16*, 480-485.

Chen, L.-L., and Carmichael, G.G. (2008). Gene regulation by SINES and inosines: biological consequences of A-to-I editing of Alu element inverted repeats. Cell Cycle *7*, 3294-3301.

Chen, L.-L., DeCerbo, J.N., and Carmichael, G.G. (2008). Alu element-mediated gene silencing. EMBO J *27*, 1694-1705.

Chen, S.H., Habib, G., Yang, C.Y., Gu, Z.W., Lee, B.R., Weng, S.A., Silberman, S.R., Cai, S.J., Deslypere, J.P., Rosseneu, M., *et al.* (1987). Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. Science *238*, 363-366.

Chen, S.H., Li, X.X., Liao, W.S., Wu, J.H., and Chan, L. (1990). RNA editing of apolipoprotein B mRNA. Sequence specificity determined by in vitro coupled transcription editing. J Biol Chem *265*, 6811-6816.

Chepelev, I., Wei, G., Tang, Q., and Zhao, K. (2009). Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Res.

Chester, A., Somasekaram, A., Tzimina, M., Jarmuz, A., Gisbourne, J., O'Keefe, R., Scott, J., and Navaratnam, N. (2003). The apolipoprotein B mRNA editing complex performs a multifunctional cycle and suppresses nonsense-mediated decay. EMBO J *22*, 3971-3982.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479-486.

Chilibeck, K.A., Wu, T., Liang, C., Schellenberg, M.J., Gesner, E.M., Lynch, J.M., and MacMillan, A.M. (2006). FRET analysis of in vivo dimerization by RNA-editing enzymes. J Biol Chem *281*, 16530-16535.

Chiu, Y.-L., Soros, V.B., Kreisberg, J.F., Stopak, K., Yonemoto, W., and Greene, W.C. (2005). Cellular APOBEC3G restricts HIV-1 infection in resting CD4+ T cells. Nature *435*, 108-114.

Cho, D.-S.C., Yang, W., Lee, J.T., Shiekhattar, R., Murray, J.M., and Nishikura, K. (2003). Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA. J Biol Chem *278*, 17093-17102.

Collins, D.W., and Jukes, T.H. (1994). Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics *20*, 386-396.

Consortium, I.H.G.S. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

Conticello, S.G. (2008). The AID/APOBEC family of nucleic acid mutators. Genome Biol *9*, 229.

Conticello, S.G., Harris, R.S., and Neuberger, M.S. (2003). The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G. Curr Biol *13*, 2009-2013.

Conticello, S.G., Langlois, M.-A., and Neuberger, M.S. (2007). Insights into DNA deaminases. Nat Struct Mol Biol *14*, 7-9.

Conticello, S.G., Thomas, C.J.F., Petersen-Mahrt, S.K., and Neuberger, M.S. (2005). Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. Mol Biol Evol *22*, 367-377.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat Rev Genet *10*, 691-703.

Crick, F.H. (1966). Codon--anticodon pairing: the wobble hypothesis. J Mol Biol *19*, 548-555.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

Dance, G.S.C., Sowden, M.P., Cartegni, L., Cooper, E., Krainer, A.R., and Smith, H.C. (2002). Two proteins essential for apolipoprotein B mRNA editing are expressed from a single gene through alternative splicing. J Biol Chem *277*, 12703-12709.

Davidson, N.O. (2002). The challenge of target sequence specificity in C-->U RNA editing. J Clin Invest *109*, 291-294.

Dawson, T.R., Sansam, C.L., and Emeson, R.B. (2004). Structure and sequence determinants required for the RNA editing of ADAR2 substrates. J Biol Chem *279*, 4941-4951.

Dickerson, S.K., Market, E., Besmer, E., and Papavasiliou, F.N. (2003). AID mediates hypermutation by deaminating single stranded DNA. J Exp Med *197*, 1291-1296.

Doehle, B.P., Schafer, A., and Cullen, B.R. (2005). Human APOBEC3B is a potent inhibitor of HIV-1 infectivity and is resistant to HIV-1 Vif. Virology *339*, 281-288.

Doria, M., Neri, F., Gallo, A., Farace, M.G., and Michienzi, A. (2009). Editing of HIV-1 RNA by the double-stranded RNA deaminase ADAR1 stimulates viral infection. Nucleic Acids Res *37*, 5848-5858.

Driscoll, D.M., Lakhe-Reddy, S., Oleksa, L.M., and Martinez, D. (1993). Induction of RNA editing at heterologous sites by sequences in apolipoprotein B mRNA. Mol Cell Biol *13*, 7288-7294.

Dunnick, W.A., Collins, J.T., Shi, J., Westfield, G., Fontaine, C., Hakimpour, P., and Papavasiliou, F.N. (2009). Switch recombination and somatic hypermutation are controlled by the heavy chain 3' enhancer region. J Exp Med *206*, 2613-2623.

Duret, L., Dorkeld, F., and Gautier, C. (1993). Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. Nucleic Acids Res *21*, 2315-2322.

Ehrenstein, M.R., and Neuberger, M.S. (1999). Deficiency in Msh2 affects the efficiency and local sequence specificity of immunoglobulin class-switch recombination: parallels with somatic hypermutation. EMBO J *18*, 3484-3490.

Esnault, C., Heidmann, O., Delebecque, F., Dewannieux, M., Ribet, D., Hance, A.J., Heidmann, T., and Schwartz, O. (2005). APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. Nature *433*, 430-433.

Farese, R.V., Ruland, S.L., Flynn, L.M., Stokowski, R.P., and Young, S.G. (1995). Knockout of the mouse apolipoprotein B gene results in embryonic lethality in homozygotes and protection against diet-induced hypercholesterolemia in heterozygotes. Proc Natl Acad Sci USA *92*, 1774-1778.

Feagin, J.E., Jasmer, D.P., and Stuart, K. (1987). Developmentally regulated addition of nucleotides within apocytochrome b transcripts in Trypanosoma brucei. Cell *49*, 337-345.

Feagin, J.E., Shaw, J.M., Simpson, L., and Stuart, K. (1988). Creation of AUG initiation codons by addition of uridines within cytochrome b transcripts of kinetoplastids. Proc Natl Acad Sci USA *85*, 539-543.

Funahashi, T., Giannoni, F., DePaoli, A.M., Skarosi, S.F., and Davidson, N.O. (1995). Tissue-specific, developmental and nutritional regulation of the gene encoding the catalytic subunit of the rat apolipoprotein B mRNA editing enzyme: functional role in the modulation of apoB mRNA editing. J Lipid Res *36*, 414-428.

Gabuzda, D.H., Lawrence, K., Langhoff, E., Terwilliger, E., Dorfman, T., Haseltine, W.A., and Sodroski, J. (1992). Role of vif in replication of human immunodeficiency virus type 1 in CD4+ T lymphocytes. J Virol *66*, 6489-6495.

Gallo, A., Keegan, L.P., Ring, G.M., and O'Connell, M.A. (2003). An ADAR that edits transcripts encoding ion channel subunits functions as a dimer. EMBO J *22*, 3421-3430.

Ganem, D. (1996). *Hepadnaviridae* and their replication. In Fields Virology, B. Fields, D. Knipe, and P. Howley, eds. (Philadelphia, Lippincott-Raven), pp. 2703-2737.

Gerber, A., Grosjean, H., Melcher, T., and Keller, W. (1998). Tad1p, a yeast tRNA-specific adenosine deaminase, is related to the mammalian pre-mRNA editing enzymes ADAR1 and ADAR2. EMBO J *17*, 4780-4789.

Gerber, A.P., and Keller, W. (1999). An adenosine deaminase that generates inosine at the wobble position of tRNAs. Science *286*, 1146-1149.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M.*, et al.* (1996). Life with 6000 genes. Science *274*, 546, 563-547.

Gommans, W.M., Mullen, S.P., and Maas, S. (2009). RNA editing: a driving force for adaptive evolution? Bioessays *31*, 1137-1145.

Green, P., Ewing, B., Miller, W., Thomas, P.J., Program, N.C.S., and Green, E.D. (2003). Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet *33*, 514-517.

Greeve, J., Altkemper, I., Dieterich, J.H., Greten, H., and Windler, E. (1993). Apolipoprotein B mRNA editing in 12 different mammalian species: hepatic expression is reflected in low concentrations of apoB-containing plasma lipoproteins. J Lipid Res *34*, 1367-1383.

Grimson, A., Farh, K.K.-H., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell *27*, 91-105.

Haché, G., Liddament, M.T., and Harris, R.S. (2005). The retroviral hypermutation specificity of APOBEC3F and APOBEC3G is governed by the C-terminal DNA cytosine deaminase domain. J Biol Chem *280*, 10920-10924.

Hao, S., and Baltimore, D. (2009). The stability of mRNA influences the temporal order of the induction of genes encoding inflammatory molecules. Nat Immunol.

Harris, R.S., Bishop, K.N., Sheehy, A.M., Craig, H.M., Petersen-Mahrt, S.K., Watt, I.N., Neuberger, M.S., and Malim, M.H. (2003). DNA deamination mediates innate immunity to retroviral infection. Cell *113*, 803-809.

Hartner, J.C., Schmittwolf, C., Kispert, A., Müller, A.M., Higuchi, M., and Seeburg, P.H. (2004). Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1. J Biol Chem *279*, 4894-4902.

Heap, G.A., Yang, J.H.M., Downes, K., Healy, B.C., Hunt, K.A., Bockett, N., Franke, L., Dubois, P.C., Mein, C.A., Dobson, R.J., *et al.* (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Hum Mol Genet *19*, 122-134.

Herb, A., Higuchi, M., Sprengel, R., and Seeburg, P.H. (1996). Q/R site editing in kainate receptor GluR5 and GluR6 pre-mRNAs requires distant intronic sequences. Proc Natl Acad Sci USA *93*, 1875-1880.

Herbert, A., Lowenhaupt, K., Spitzner, J., and Rich, A. (1995). Chicken double-stranded RNA adenosine deaminase has apparent specificity for Z-DNA. Proc Natl Acad Sci USA *92*, 7550-7554.

Hersberger, M., Patarroyo-White, S., Arnold, K.S., and Innerarity, T.L. (1999). Phylogenetic analysis of the apolipoprotein B mRNA-editing region. Evidence for a secondary structure between the mooring sequence and the 3' efficiency element. J Biol Chem *274*, 34590-34597.

Higuchi, M., Maas, S., Single, F.N., Hartner, J., Rozov, A., Burnashev, N., Feldmeyer, D., Sprengel, R., and Seeburg, P.H. (2000). Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. Nature *406*, 78-81.

Higuchi, M., Single, F.N., Köhler, M., Sommer, B., Sprengel, R., and Seeburg, P.H. (1993). RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. Cell *75*, 1361-1370.

Hirano, K., Min, J., Funahashi, T., and Davidson, N.O. (1997). Cloning and characterization of the rat apobec-1 gene: a comparative analysis of gene structure and promoter usage in rat and mouse. J Lipid Res *38*, 1103-1119.

Hirano, K., Young, S.G., Farese, R.V., Ng, J., Sande, E., Warburton, C., Powell-Braxton, L.M., and Davidson, N.O. (1996). Targeted disruption of the mouse apobec-1 gene abolishes apolipoprotein B mRNA editing and eliminates apolipoprotein B48. J Biol Chem *271*, 9887-9890.

Hundley, H.A., and Bass, B.L. (2010). ADAR editing in double-stranded UTRs and other noncoding RNA sequences. Trends in biochemical sciences.

Illumina (2008a). Using SBS Sequencing Kit v3 on the Genome Analyzer (San Diego, CA).

Illumina (2008b). Using the Single-Read Cluster Generation Kit v2 on the Cluster Station (San DIego, CA).

Imai, K., Slupphaug, G., Lee, W.-I., Revy, P., Nonoyama, S., Catalan, N., Yel, L., Forveille, M., Kavli, B., Krokan, H.E., *et al.* (2003). Human uracil-DNA glycosylase deficiency associated with profoundly impaired immunoglobulin class-switch recombination. Nat Immunol *4*, 1023-1028.

Ireton, G.C., McDermott, G., Black, M.E., and Stoddard, B.L. (2002). The structure of Escherichia coli cytosine deaminase. J Mol Biol *315*, 687-697.

Ito, S., Nagaoka, H., Shinkura, R., Begum, N., Muramatsu, M., Nakata, M., and Honjo, T. (2004). Activation-induced cytidine deaminase shuttles between nucleus and cytoplasm like apolipoprotein B mRNA editing catalytic polypeptide 1. Proc Natl Acad Sci USA *101*, 1975-1980.

Jackson, R.J., and Standart, N. (2007). How do microRNAs regulate gene expression? Sci STKE *2007*, re1.

Jaikaran, D.C.J., Collins, C.H., and MacMillan, A.M. (2002). Adenosine to inosine editing by ADAR2 requires formation of a ternary complex on the GluR-B R/G site. J Biol Chem *277*, 37624-37629.

Jambhekar, A., and Derisi, J.L. (2007). Cis-acting determinants of asymmetric, cytoplasmic RNA transport. RNA *13*, 625-642.

Jansen, R.P. (2001). mRNA localization: message on the move. Nat Rev Mol Cell Biol *2*, 247-256.

Jarmuz, A., Chester, A., Bayliss, J., Gisbourne, J., Dunham, I., Scott, J., and Navaratnam, N. (2002). An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. Genomics *79*, 285-296.

Jin, Y., Zhang, W., and Li, Q. (2009). Origins and evolution of ADAR-mediated RNA editing. IUBMB Life *61*, 572-578.

Johansson, E., Mejlhede, N., Neuhard, J., and Larsen, S. (2002). Crystal structure of the tetrameric cytidine deaminase from Bacillus subtilis at 2.0 A resolution. Biochemistry *41*, 2563-2570.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., *et al.* (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res *14*, 331-342.

Kawahara, Y., and Nishikura, K. (2006). Extensive adenosine-to-inosine editing detected in Alu repeats of antisense RNAs reveals scarcity of sense-antisense duplex formation. FEBS Lett *580*, 2301-2305.

Kawahara, Y., Zinshteyn, B., Chendrimada, T.P., Shiekhattar, R., and Nishikura, K. (2007). RNA editing of the microRNA-151 precursor blocks cleavage by the Dicer-TRBP complex. EMBO Rep *8*, 763-769.

Kendrick, J.S., Chan, L., and Higgins, J.A. (2001). Superior role of apolipoprotein B48 over apolipoprotein B100 in chylomicron assembly and fat absorption: an investigation of apobec-1 knock-out and wild-type mice. Biochem J *356*, 821-827.

Khan, M.A., Kao, S., Miyagi, E., Takeuchi, H., Goila-Gaur, R., Opi, S., Gipson, C.L., Parslow, T.G., Ly, H., and Strebel, K. (2005). Viral RNA is required for the association of APOBEC3G with human immunodeficiency virus type 1 nucleoprotein complexes. J Virol *79*, 5870-5874.

Kim, D.D.Y., Kim, T.T.Y., Walsh, T., Kobayashi, Y., Matise, T.C., Buyske, S., and Gabriel, A. (2004). Widespread RNA editing of embedded alu elements in the human transcriptome. Genome Res *14*, 1719-1725.

Kim, U., Wang, Y., Sanford, T., Zeng, Y., and Nishikura, K. (1994). Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. Proc Natl Acad Sci USA *91*, 11457-11461.

Kobayashi, M., Takaori-Kondo, A., Miyauchi, Y., Iwai, K., and Uchiyama, T. (2005). Ubiquitination of APOBEC3G by an HIV-1 Vif-Cullin5-Elongin B-Elongin C complex is essential for Vif function. J Biol Chem *280*, 18573-18578.

Kumar, M., and Carmichael, G.G. (1997). Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. Proc Natl Acad Sci USA *94*, 3542-3547.

Kuo, M.Y., Chao, M., and Taylor, J. (1989). Initiation of replication of the human hepatitis delta virus genome from cloned DNA: role of delta antigen. J Virol *63*, 1945-1950.

Landweber, L.F., and Gilbert, W. (1993). RNA editing as a source of genetic variation. Nature *363*, 179-182.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Larocque, D., Pilotte, J., Chen, T., Cloutier, F., Massie, B., Pedraza, L., Couture, R., Lasko, P., Almazan, G., and Richard, S. (2002). Nuclear retention of MBP mRNAs in the quaking viable mice. Neuron *36*, 815-829.

Lau, P.P., Chang, B.H., and Chan, L. (2001a). Two-hybrid cloning identifies an RNA-binding protein, GRY-RBP, as a component of apobec-1 editosome. Biochem Biophys Res Commun *282*, 977-983.

Lau, P.P., Villanueva, H., Kobayashi, K., Nakamuta, M., Chang, B.H., and Chan, L. (2001b). A DnaJ protein, apobec-1-binding protein-2, modulates apolipoprotein B mRNA editing. J Biol Chem *276*, 46445-46452.

Lau, P.P., Xiong, W.J., Zhu, H.J., Chen, S.H., and Chan, L. (1991). Apolipoprotein B mRNA editing is an intranuclear event that occurs posttranscriptionally coincident with splicing and polyadenylation. J Biol Chem *266*, 20550-20554.

Lau, P.P., Zhu, H.J., Baldini, A., Charnsangavej, C., and Chan, L. (1994). Dimeric structure of a human apolipoprotein B mRNA editing protein and cloning and chromosomal localization of its gene. Proc Natl Acad Sci USA *91*, 8522-8526.

Lau, P.P., Zhu, H.J., Nakamuta, M., and Chan, L. (1997). Cloning of an Apobec-1-binding protein that also interacts with apolipoprotein B mRNA and evidence for its involvement in RNA editing. J Biol Chem *272*, 1452-1455.

Lecossier, D., Bouchonnet, F., Clavel, F., and Hance, A.J. (2003). Hypermutation of HIV-1 DNA in the absence of the Vif protein. Science *300*, 1112.

Lee, W.-I., Torgerson, T.R., Schumacher, M.J., Yel, L., Zhu, Q., and Ochs, H.D. (2005). Molecular analysis of a large cohort of patients with the hyper immunoglobulin M (IgM) syndrome. Blood *105*, 1881-1890.

Lehmann, D.M., Galloway, C.A., MacElrevey, C., Sowden, M.P., Wedekind, J.E., and Smith, H.C. (2007). Functional characterization of APOBEC-1 complementation factor phosphorylation sites. Biochim Biophys Acta *1773*, 408-418.

Lehmann, D.M., Galloway, C.A., Sowden, M.P., and Smith, H.C. (2006). Metabolic regulation of apoB mRNA editing is associated with phosphorylation of APOBEC-1 complementation factor. Nucleic Acids Res *34*, 3299-3308.

Lehmann, K.A., and Bass, B.L. (1999). The importance of internal loops within RNA substrates of ADAR1. J Mol Biol *291*, 1-13.

Lehmann, K.A., and Bass, B.L. (2000). Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry *39*, 12875-12884.

Lellek, H., Kirsten, R., Diehl, I., Apostel, F., Buck, F., and Greeve, J. (2000). Purification and molecular cloning of a novel essential component of the apolipoprotein B mRNA editing enzyme-complex. J Biol Chem *275*, 19848-19856.

Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D., *et al.* (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol *22*, 1001-1005.

Levanon, E.Y., Hallegger, M., Kinar, Y., Shemesh, R., Djinovic-Carugo, K., Rechavi, G., Jantsch, M.F., and Eisenberg, E. (2005). Evolutionarily conserved human targets of adenosine to inosine RNA editing. Nucleic Acids Res *33*, 1162-1168.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, G.P.D.P. (2009a). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res *18*, 1851-1858.

Li, J.B., Levanon, E.Y., Yoon, J.-K., Aach, J., Xie, B., Leproust, E., Zhang, K., Gao, Y., and Church, G.M. (2009b). Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science *324*, 1210-1213.

Liang, H., and Landweber, L.F. (2007). Hypothesis: RNA editing of microRNA target sites in humans? RNA *13*, 463-467.

Liao, W., Hong, S.H., Chan, B.H., Rudolph, F.B., Clark, S.C., and Chan, L. (1999). APOBEC-2, a cardiac- and skeletal muscle-specific member of the cytidine deaminase supergene family. Biochem Biophys Res Commun *260*, 398-404.

Limbach, P.A., Crain, P.F., and McCloskey, J.A. (1994). Summary: the modified nucleosides of RNA. Nucleic Acids Res *22*, 2183-2196.

Lipman, D.J. (1997). Making (anti)sense of non-coding sequence conservation. Nucleic Acids Res *25*, 3580-3583.

Liu, M., Duke, J.L., Richter, D.J., Vinuesa, C.G., Goodnow, C.C., Kleinstein, S.H., and Schatz, D.G. (2008). Two levels of protection for the B cell genome during somatic hypermutation. Nature *451*, 841-845.

Lomeli, H., Mosbacher, J., Melcher, T., Höger, T., Geiger, J.R., Kuner, T., Monyer, H., Higuchi, M., Bach, A., and Seeburg, P.H. (1994). Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. Science *266*, 1709-1713.

Low, A., Okeoma, C.M., Lovsin, N., de las Heras, M., Taylor, T.H., Peterlin, B.M., Ross, S.R., and Fan, H. (2009). Enhanced replication and pathogenesis of Moloney murine leukemia virus in mice defective in the murine APOBEC3 gene. Virology *385*, 455-463.

Luciano, D.J., Mirsky, H., Vendetti, N.J., and Maas, S. (2004). RNA editing of a miRNA precursor. RNA *10*, 1174-1177.

Luo, G.X., Chao, M., Hsieh, S.Y., Sureau, C., Nishikura, K., and Taylor, J. (1990). A specific base transition occurs on replicating hepatitis delta virus RNA. J Virol *64*, 1021-1027.

Maas, S., Gerber, A.P., and Rich, A. (1999). Identification and characterization of a human tRNA-specific adenosine deaminase related to the ADAR family of pre-mRNA editing enzymes. Proc Natl Acad Sci USA *96*, 8895-8900.

MacGinnitie, A.J., Anant, S., and Davidson, N.O. (1995). Mutagenesis of apobec-1, the catalytic subunit of the mammalian apolipoprotein B mRNA editing enzyme, reveals distinct domains that mediate cytosine nucleoside deaminase, RNA binding, and RNA editing activity. J Biol Chem *270*, 14768-14775.

Madani, N., and Kabat, D. (1998). An endogenous inhibitor of human immunodeficiency virus in human lymphocytes is overcome by the viral Vif protein. J Virol *72*, 10251-10255.

Malim, M.H., and Emerman, M. (2008). HIV-1 accessory proteins--ensuring viral survival in a hostile environment. Cell Host Microbe *3*, 388-398.

Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., and Trono, D. (2003). Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. Nature *424*, 99-103.

Mangeat, B., Turelli, P., Liao, S., and Trono, D. (2004). A single amino acid determinant governs the species-specific sensitivity of APOBEC3G to Vif action. J Biol Chem *279*, 14481-14483.

Marass, F., and Upton, C. (2009). Sequence Searcher: A Java tool to perform regular expression and fuzzy searches of multiple DNA and protein sequences. BMC Res Notes *2*, 14.

Mariani, R., Chen, D., Schröfelbauer, B., Navarro, F., König, R., Bollman, B., Münk, C., Nymark-McMahon, H., and Landau, N.R. (2003). Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. Cell *114*, 21-31.

Marin, M., Rose, K.M., Kozak, S.L., and Kabat, D. (2003). HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. Nat Med *9*, 1398-1403.

Maris, C., and Allain, F.H.-T. (2009). Structure of RNA Editing Substrates and Their Recognition by RNA Base Deaminase. In DNA and RNA Modification Enzymes: Structure, Mechanism, Function and Evolution, H. Grosjean, ed. (Austin, TX, USA, Landes Bioscience), pp. 224-242.

Maris, C., Masse, J., Chester, A., Navaratnam, N., and Allain, F.H.-T. (2005). NMR structure of the apoB mRNA stem-loop and its interaction with the C to U editing APOBEC1 complementary factor. RNA *11*, 173-186.

Martin, A., Bardwell, P.D., Woo, C.J., Fan, M., Shulman, M.J., and Scharff, M.D. (2002). Activation-induced cytidine deaminase turns on somatic hypermutation in hybridomas. Nature *415*, 802-806.

Martomo, S.A., Yang, W.W., and Gearhart, P.J. (2004). A role for Msh6 but not Msh3 in somatic hypermutation and class switch recombination. J Exp Med *200*, 61-68.

Martomo, S.A., Yang, W.W., Wersto, R.P., Ohkumo, T., Kondo, Y., Yokoi, M., Masutani, C., Hanaoka, F., and Gearhart, P.J. (2005). Different mutation signatures in DNA polymerase eta- and MSH6-deficient mice suggest separate roles in antibody diversification. Proc Natl Acad Sci USA *102*, 8656-8661.

Mehle, A., Strack, B., Ancuta, P., Zhang, C., McPike, M., and Gabuzda, D. (2004). Vif overcomes the innate antiviral activity of APOBEC3G by promoting its degradation in the ubiquitin-proteasome pathway. J Biol Chem *279*, 7792-7798.

Mehta, A., Banerjee, S., and Driscoll, D.M. (1996). Apobec-1 interacts with a 65-kDa complementing protein to edit apolipoprotein-B mRNA in vitro. J Biol Chem *271*, 28294-28299.

Mehta, A., Kinter, M.T., Sherman, N.E., and Driscoll, D.M. (2000). Molecular cloning of apobec-1 complementation factor, a novel RNA-binding protein involved in the editing of apolipoprotein B mRNA. Mol Cell Biol *20*, 1846-1854.

Melcher, T., Maas, S., Herb, A., Sprengel, R., Higuchi, M., and Seeburg, P.H. (1996a). RED2, a brain-specific member of the RNA-specific adenosine deaminase family. J Biol Chem *271*, 31795-31798.

Melcher, T., Maas, S., Herb, A., Sprengel, R., Seeburg, P.H., and Higuchi, M. (1996b). A mammalian RNA editing enzyme. Nature *379*, 460-464.

Mewes, H.W., Albermann, K., Bähr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S.G., *et al.* (1997). Overview of the yeast genome. Nature *387*, 7-65.

Mian, I.S., Moser, M.J., Holley, W.R., and Chatterjee, A. (1998). Statistical modelling and phylogenetic analysis of a deaminase domain. J Comput Biol *5*, 57-72.

Mikl, M.C., Watt, I.N., Lu, M., Reik, W., Davies, S.L., Neuberger, M.S., and Rada, C. (2005). Mice deficient in APOBEC2 and APOBEC3. Mol Cell Biol *25*, 7270-7277.

Misra, S., Crosby, M.A., Mungall, C.J., Matthews, B.B., Campbell, K.S., Hradecky, P., Huang, Y., Kaminker, J.S., Millburn, G.H., Prochnik, S.E., *et al.* (2002). Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. Genome Biol *3*, RESEARCH0083.

Morrison, J.R., Pászty, C., Stevens, M.E., Hughes, S.D., Forte, T., Scott, J., and Rubin, E.M. (1996). Apolipoprotein B RNA editing enzyme-deficient mice are viable despite alterations in lipoprotein metabolism. Proc Natl Acad Sci USA *93*, 7154-7159.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621-628.

Muckenfuss, H., Hamdorf, M., Held, U., Perkovic, M., Löwer, J., Cichutek, K., Flory, E., Schumann, G.G., and Münk, C. (2006). APOBEC3 proteins inhibit human LINE-1 retrotransposition. J Biol Chem *281*, 22161-22172.

Mukhopadhyay, D., Anant, S., Lee, R.M., Kennedy, S., Viskochil, D., and Davidson, N.O. (2002). C-->U editing of neurofibromatosis 1 mRNA occurs in tumors that express both the type II transcript and apobec-1, the catalytic subunit of the apolipoprotein B mRNA-editing enzyme. Am J Hum Genet *70*, 38-50.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell *102*, 553-563.

Muramatsu, M., Sankaranand, V.S., Anant, S., Sugai, M., Kinoshita, K., Davidson, N.O., and Honjo, T. (1999). Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. J Biol Chem *274*, 18470-18476.

Murphy, D.G., Dimock, K., and Kang, C.Y. (1991). Numerous transitions in human parainfluenza virus 3 RNA recovered from persistently infected cells. Virology *181*, 760-763.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344-1349.

Nakamuta, M., Oka, K., Krushkal, J., Kobayashi, K., Yamamoto, M., Li, W.H., and Chan, L. (1995). Alternative mRNA splicing and differential promoter utilization determine tissue-specific expression of the apolipoprotein B mRNA-editing protein (Apobec1) gene in mice. Structure and evolution of Apobec1 and related nucleoside/nucleotide deaminases. J Biol Chem *270*, 13042-13056.

Navaratnam, N., Bhattacharya, S., Fujino, T., Patel, D., Jarmuz, A.L., and Scott, J. (1995). Evolutionary origins of apoB mRNA editing: catalysis by a cytidine deaminase that has acquired a novel RNA-binding motif at its active site. Cell *81*, 187-195.

Nishikura, K., Yoo, C., Kim, U., Murray, J.M., Estes, P.A., Cash, F.E., and Liebhaber, S.A. (1991). Substrate specificity of the dsRNA unwinding/modifying activity. EMBO J *10*, 3523-3532.

Niswender, C.M., Copeland, S.C., Herrick-Davis, K., Emeson, R.B., and Sanders-Bush, E. (1999). RNA editing of the human serotonin 5-hydroxytryptamine 2C receptor silences constitutive activity. J Biol Chem *274*, 9472-9478.

Niswender, C.M., Sanders-Bush, E., and Emeson, R.B. (1998). Identification and characterization of RNA editing events within the 5-HT2C receptor. Ann N Y Acad Sci *861*, 38-48.

O'Connell, M.A., Krause, S., Higuchi, M., Hsuan, J.J., Totty, N.F., Jenny, A., and Keller, W. (1995). Cloning of cDNAs encoding mammalian double-stranded RNA-specific adenosine deaminase. Mol Cell Biol *15*, 1389-1397.

O'Hara, P.J., Nichol, S.T., Horodyski, F.M., and Holland, J.J. (1984). Vesicular stomatitis virus defective interfering particles can contain extensive genomic sequence rearrangements and base substitutions. Cell *36*, 915-924.

Ohman, M., Källman, A.M., and Bass, B.L. (2000). In vitro analysis of the binding of ADAR2 to the pre-mRNA encoding the GluR-B R/G site. RNA *6*, 687-697.

Oka, K., Kobayashi, K., Sullivan, M., Martinez, J., Teng, B.B., Ishimura-Oka, K., and Chan, L. (1997). Tissue-specific inhibition of apolipoprotein B mRNA editing in the liver by adenovirus-mediated transfer of a dominant negative mutant APOBEC-1 leads to increased low density lipoprotein in mice. J Biol Chem *272*, 1456-1460.

Okazaki, I.-M., Kinoshita, K., Muramatsu, M., Yoshikawa, K., and Honjo, T. (2002). The AID enzyme induces class switch recombination in fibroblasts. Nature *416*, 340-345.

Okeoma, C.M., Low, A., Bailis, W., Fan, H.Y., Peterlin, B.M., and Ross, S.R. (2009). Induction of APOBEC3 in vivo causes increased restriction of retrovirus infection. J Virol *83*, 3486-3495.

Osenberg, S., Dominissini, D., Rechavi, G., and Eisenberg, E. (2009). Widespread cleavage of A-to-I hyperediting substrates. RNA *15*, 1632-1639.

Palladino, M.J., Keegan, L.P., O'Connell, M.A., and Reenan, R.A. (2000a). A-to-I pre-mRNA editing in Drosophila is primarily involved in adult nervous system function and integrity. Cell *102*, 437-449.

Palladino, M.J., Keegan, L.P., O'Connell, M.A., and Reenan, R.A. (2000b). dADAR, a Drosophila double-stranded RNA-specific adenosine deaminase is highly developmentally regulated and is itself a target for RNA editing. RNA *6*, 1004-1018.

Pan, X., and Weissman, S.M. (2002). An approach for global scanning of single nucleotide variations. Proc Natl Acad Sci USA *99*, 9346-9351.

Panigrahi, A.K., Allen, T.E., Stuart, K., Haynes, P.A., and Gygi, S.P. (2003a). Mass spectrometric analysis of the editosome and other multiprotein complexes in Trypanosoma brucei. J Am Soc Mass Spectrom *14*, 728-735.

Panigrahi, A.K., Schnaufer, A., Ernst, N.L., Wang, B., Carmean, N., Salavati, R., and Stuart, K. (2003b). Identification of novel components of Trypanosoma brucei editosomes. RNA *9*, 484-492.

Patenaude, A.-M., Orthwein, A., Hu, Y., Campo, V.A., Kavli, B., Buschiazzo, A., and Di Noia, J.M. (2009). Active nuclear import and cytoplasmic retention of activation-induced deaminase. Nat Struct Mol Biol *16*, 517-527.

Pham, P., Bransteitter, R., Petruska, J., and Goodman, M.F. (2003). Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature *424*, 103-107.

Phuphuakrat, A., Kraiwong, R., Boonarkart, C., Lauhakirti, D., Lee, T.-H., and Auewarakul, P. (2008). Double-stranded RNA adenosine deaminases enhance expression of human immunodeficiency virus type 1 proteins. J Virol *82*, 10864-10872.

Polson, A.G., and Bass, B.L. (1994). Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. EMBO J *13*, 5701-5711.

Polson, A.G., Bass, B.L., and Casey, J.L. (1996). RNA editing of hepatitis delta virus antigenome by dsRNA-adenosine deaminase. Nature *380*, 454-456.

Powell, L.M., Wallis, S.C., Pease, R.J., Edwards, Y.H., Knott, T.J., and Scott, J. (1987). A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell *50*, 831-840.

Prochnow, C., Bransteitter, R., Klein, M.G., Goodman, M.F., and Chen, X.S. (2007). The APOBEC-2 crystal structure and functional implications for the deaminase AID. Nature *445*, 447-451.

Rada, C., Ehrenstein, M.R., Neuberger, M.S., and Milstein, C. (1998). Hot spot focusing of somatic hypermutation in MSH2-deficient mice suggests two stages of mutational targeting. Immunity *9*, 135-141.

Rada, C., Williams, G.T., Nilsen, H., Barnes, D.E., Lindahl, T., and Neuberger, M.S. (2002). Immunoglobulin isotype switching is inhibited and somatic hypermutation perturbed in UNG-deficient mice. Curr Biol *12*, 1748-1755.

Raghavan, S.C., Swanson, P.C., Wu, X., Hsieh, C.-L., and Lieber, M.R. (2004). A non-B-DNA structure at the Bcl-2 major breakpoint region is cleaved by the RAG complex. Nature *428*, 88-93.

Ramiro, A.R., Jankovic, M., Callen, E., Difilippantonio, S., Chen, H.-T., McBride, K.M., Eisenreich, T.R., Chen, J., Dickins, R.A., Lowe, S.W., *et al.* (2006). Role of genomic instability and p53 in AID-induced c-myc-Igh translocations. Nature *440*, 105-109.

Ray, P.S., Jia, J., Yao, P., Majumder, M., Hatzoglou, M., and Fox, P.L. (2009). A stress-responsive RNA switch regulates VEGFA expression. Nature *457*, 915-919.

Rebagliati, M.R., and Melton, D.A. (1987). Antisense RNA injections in fertilized frog eggs reveal an RNA duplex unwinding activity. Cell *48*, 599-605.

Revy, P., Muto, T., Levy, Y., Geissmann, F., Plebani, A., Sanal, O., Catalan, N., Forveille, M., Dufourcq-Labelouse, R., Gennery, A., *et al.* (2000). Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). Cell *102*, 565-575.

Riedmann, E.M., Schopoff, S., Hartner, J.C., and Jantsch, M.F. (2008). Specificity of ADAR-mediated RNA editing in newly identified targets. RNA *14*, 1110-1118.

Robbiani, D.F., Bothmer, A., Callen, E., Reina-San-Martin, B., Dorsett, Y., Difilippantonio, S., Bolland, D.J., Chen, H.T., Corcoran, A.E., Nussenzweig, A., *et al.* (2008). AID is required for the chromosomal breaks in c-myc that lead to c-myc/IgH translocations. Cell *135*, 1028-1038.

Rogozin, I.B., Basu, M.K., Jordan, I.K., Pavlov, Y.I., and Koonin, E.V. (2005). APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases predicted by computational analysis. Cell Cycle *4*, 1281-1285.

Rosenberg, B.R., and Papavasiliou, F.N. (2007). Beyond SHM and CSR: AID and related cytidine deaminases in the host response to viral infection. Adv Immunol *94*, 215-244.

Rozenski, J., Crain, P.F., and McCloskey, J.A. (1999). The RNA Modification Database: 1999 update. Nucleic Acids Res *27*, 196-197.

Rubio, M.A.T., Pastar, I., Gaston, K.W., Ragone, F.L., Janzen, C.J., Cross, G.A.M., Papavasiliou, F.N., and Alfonzo, J.D. (2007). An adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U deamination of DNA. Proc Natl Acad Sci USA *104*, 7821-7826.

Ryu, W.S., Bayer, M., and Taylor, J. (1992). Assembly of hepatitis delta virus particles. J Virol *66*, 2310-2315.

Sato, Y., Probst, H.C., Tatsumi, R., Ikeuchi, Y., Neuberger, M.S., and Rada, C. (2009). Deficiency in APOBEC2 leads to a shift in muscle fiber-type, diminished body mass and myopathy. J Biol Chem.

Sawyer, S.L., Emerman, M., and Malik, H.S. (2004). Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS Biol *2*, E275.

Scadden, A.D., and Smith, C.W. (1997). A ribonuclease specific for inosine-containing RNA: a potential role in antiviral defence? EMBO J *16*, 2140-2149.

Scadden, A.D., and Smith, C.W. (2001). RNAi is antagonized by A-->I hyper-editing. EMBO Rep *2*, 1107-1111.

Schäfer, A., Bogerd, H.P., and Cullen, B.R. (2004). Specific packaging of APOBEC3G into HIV-1 virions is mediated by the nucleocapsid domain of the gag polyprotein precursor. Virology *328*, 163-168.

Schnaufer, A., Ernst, N.L., Palazzo, S.S., O'Rear, J., Salavati, R., and Stuart, K. (2003). Separate insertion and deletion subcomplexes of the Trypanosoma brucei RNA editing complex. Mol Cell *12*, 307-319.

Schnaufer, A., Panigrahi, A.K., Panicucci, B., Igo, R.P., Wirtz, E., Salavati, R., and Stuart, K. (2001). An RNA ligase essential for RNA editing and survival of the bloodstream form of Trypanosoma brucei. Science *291*, 2159-2162.

Schröfelbauer, B., Senger, T., Manning, G., and Landau, N.R. (2006). Mutational alteration of human immunodeficiency virus type 1 Vif allows for functional interaction with nonhuman primate APOBEC3G. J Virol *80*, 5984-5991.

Seiwert, S.D., and Stuart, K. (1994). RNA editing: transfer of genetic information from gRNA to precursor mRNA in vitro. Science *266*, 114-117.

Shah, R.R., Knott, T.J., Legros, J.E., Navaratnam, N., Greeve, J.C., and Scott, J. (1991). Sequence requirements for the editing of apolipoprotein B mRNA. J Biol Chem *266*, 16301-16304.

Shaw, J.M., Campbell, D., and Simpson, L. (1989). Internal frameshifts within the mitochondrial genes for cytochrome oxidase subunit II and maxicircle unidentified reading frame 3 of Leishmania tarentolae are corrected by RNA

editing: evidence for translation of the edited cytochrome oxidase subunit II mRNA. Proc Natl Acad Sci USA *86*, 6220-6224.

Shaw, J.M., Feagin, J.E., Stuart, K., and Simpson, L. (1988). Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. Cell *53*, 401-411.

Sheehy, A.M., Gaddis, N.C., Choi, J.D., and Malim, M.H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature *418*, 646-650.

Sheehy, A.M., Gaddis, N.C., and Malim, M.H. (2003). The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. Nat Med *9*, 1404-1407.

Shivarov, V., Shinkura, R., and Honjo, T. (2008). Dissociation of in vitro DNA deamination activity and physiological functions of AID mutants. Proc Natl Acad Sci USA *105*, 15866-15871.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034-1050.

Simon, J.H., Gaddis, N.C., Fouchier, R.A., and Malim, M.H. (1998). Evidence for a newly discovered cellular anti-HIV-1 phenotype. Nat Med *4*, 1397-1400.

Simon, V., Zennou, V., Murray, D., Huang, Y., Ho, D.D., and Bieniasz, P.D. (2005). Natural variation in Vif: differential impact on APOBEC3G/3F and a potential role in HIV-1 diversification. PLoS Pathog *1*, e6.

Simpson, L., Aphasizhev, R., Gao, G., and Kang, X. (2004). Mitochondrial proteins and complexes in Leishmania and Trypanosoma involved in U-insertion/deletion RNA editing. RNA *10*, 159-170.

Simpson, L., Neckelmann, N., de la Cruz, V.F., Simpson, A.M., Feagin, J.E., Jasmer, D.P., and Stuart, J.E. (1987). Comparison of the maxicircle (mitochondrial) genomes of Leishmania tarentolae and Trypanosoma brucei at the level of nucleotide sequence. J Biol Chem *262*, 6182-6196.

Skuse, G.R., Cappione, A.J., Sowden, M., Metheny, L.J., and Smith, H.C. (1996). The neurofibromatosis type I messenger RNA undergoes base-modification RNA editing. Nucleic Acids Res *24*, 478-485.

Slavov, D., Clark, M., and Gardiner, K. (2000a). Comparative analysis of the RED1 and RED2 A-to-I RNA editing genes from mammals, pufferfish and zebrafish. Gene *250*, 41-51.

Slavov, D., Crnogorac-Jurcević, T., Clark, M., and Gardiner, K. (2000b). Comparative analysis of the DRADA A-to-I RNA editing gene from mammals, pufferfish and zebrafish. Gene *250*, 53-60.

Smith, H.C. (2009). The APOBEC1 Paradigm for Mammalian Cytidine Deaminases That Edit DNA and RNA. In DNA and RNA Modification Enzymes: Structure, Mechanism, Function and Evolution, H. Grosjean, ed. (Austin, TX, USA, Landes Bioscience), pp. 181-202.

Smith, H.C., Wedekind, J.E., Xie, K., and Sowden, M.P. (2005). Mammalian C to U editing. Topics in Current Genetics *12/2005*.

Sommer, B., Köhler, M., Sprengel, R., and Seeburg, P.H. (1991). RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. Cell *67*, 11-19.

Soros, V.B., Yonemoto, W., and Greene, W.C. (2007). Newly synthesized APOBEC3G is incorporated into HIV virions, inhibited by HIV RNA, and subsequently activated by RNase H. PLoS Pathog *3*, e15.

Sowden, M., Hamm, J.K., and Smith, H.C. (1996a). Overexpression of APOBEC-1 results in mooring sequence-dependent promiscuous RNA editing. J Biol Chem *271*, 3011-3017.

Sowden, M., Hamm, J.K., Spinelli, S., and Smith, H.C. (1996b). Determinants involved in regulating the proportion of edited apolipoprotein B RNAs. RNA *2*, 274-288.

Sowden, M.P., Eagleton, M.J., and Smith, H.C. (1998). Apolipoprotein B RNA sequence 3' of the mooring sequence and cellular sources of auxiliary factors determine the location and extent of promiscuous editing. Nucleic Acids Res *26*, 1644-1652.

Sowden, M.P., Lehmann, D.M., Lin, X., Smith, C.O., and Smith, H.C. (2004). Identification of novel alternative splice variants of APOBEC-1 complementation factor with different capacities to support apolipoprotein B mRNA editing. J Biol Chem *279*, 197-206.

Sowden, M.P., and Smith, H.C. (2001). Commitment of apolipoprotein B RNA to the splicing pathway regulates cytidine-to-uridine editing-site utilization. Biochem J *359*, 697-705.

Sprinzl, M., Steegborn, C., Hübel, F., and Steinberg, S. (1996). Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res *24*, 68-72.

Standart, N., and Jackson, R.J. (2007). MicroRNAs repress translation of m7Gppp-capped target mRNAs in vitro by inhibiting initiation and promoting deadenylation. Genes Dev *21*, 1975-1982.

Stephens, O.M., Haudenschild, B.L., and Beal, P.A. (2004). The binding selectivity of ADAR2's dsRBMs contributes to RNA-editing selectivity. Chem Biol *11*, 1239-1250.

Stopak, K., de Noronha, C., Yonemoto, W., and Greene, W.C. (2003). HIV-1 Vif blocks the antiviral activity of APOBEC3G by impairing both its translation and intracellular stability. Mol Cell *12*, 591-601.

Stuart, K., Allen, T.E., Heidmann, S., and Seiwert, S.D. (1997). RNA editing in kinetoplastid protozoa. Microbiol Mol Biol Rev *61*, 105-120.

Stuart, K., Panigrahi, A.K., and Schnaufer, A. (2004). Identification and characterization of trypanosome RNA-editing complex components. Methods Mol Biol *265*, 273-291.

Sturm, N.R., and Simpson, L. (1990a). Kinetoplast DNA minicircles encode guide RNAs for editing of cytochrome oxidase subunit III mRNA. Cell *61*, 879-884.

Sturm, N.R., and Simpson, L. (1990b). Partially edited mRNAs for cytochrome b and subunit III of cytochrome oxidase from Leishmania tarentolae mitochondria: RNA editing intermediates. Cell *61*, 871-878.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., *et al.* (2008). A global

view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science *321*, 956-960.

Svarovskaia, E.S., Xu, H., Mbisa, J.L., Barr, R., Gorelick, R.J., Ono, A., Freed, E.O., Hu, W.-S., and Pathak, V.K. (2004). Human apolipoprotein B mRNA-editing enzyme-catalytic polypeptide-like 3G (APOBEC3G) is incorporated into HIV-1 virions through interactions with viral and nonviral RNAs. J Biol Chem *279*, 35822-35828.

Taylor, J., Schenck, I., Blankenberg, D., and Nekrutenko, A. (2007). Using galaxy to perform large-scale interactive data analyses. Curr Protoc Bioinformatics *Chapter 10*, Unit 10.15.

Teng, B., Burant, C.F., and Davidson, N.O. (1993). Molecular cloning of an apolipoprotein B messenger RNA editing protein. Science *260*, 1816-1819.

Tenoever, B.R., Ng, S.-L., Chua, M.A., McWhirter, S.M., García-Sastre, A., and Maniatis, T. (2007). Multiple functions of the IKK-related kinase IKKepsilon in interferon-mediated antiviral immunity. Science *315*, 1274-1278.

Tonkin, L.A., Saccomanno, L., Morse, D.P., Brodigan, T., Krause, M., and Bass, B.L. (2002). RNA editing by ADARs is important for normal behavior in Caenorhabditis elegans. EMBO J *21*, 6025-6035.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Wagner, R.W., Smith, J.E., Cooperman, B.S., and Nishikura, K. (1989). A double-stranded RNA unwinding activity introduces structural alterations by means of adenosine to inosine conversions in mammalian cells and Xenopus eggs. Proc Natl Acad Sci USA *86*, 2647-2651.

Wakae, K., Magor, B.G., Saunders, H., Nagaoka, H., Kawamura, A., Kinoshita, K., Honjo, T., and Muramatsu, M. (2006). Evolution of class switch recombination function in fish activation-induced cytidine deaminase, AID. Int Immunol *18*, 41-47.

Wang, L., Kimble, J., and Wickens, M. (2004a). Tissue-specific modification of gld-2 mRNA in C. elegans: likely C-to-U editing. RNA *10*, 1444-1448.

Wang, Q., Miyakoda, M., Yang, W., Khillan, J., Stachura, D.L., Weiss, M.J., and Nishikura, K. (2004b). Stress-induced apoptosis associated with null mutation of ADAR1 RNA editing deaminase gene. J Biol Chem *279*, 4952-4961.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Wedekind, J.E., Dance, G.S.C., Sowden, M.P., and Smith, H.C. (2003). Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. Trends Genet *19*, 207-216.

Wolf, J., Gerber, A.P., and Keller, W. (2002). tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli. EMBO J *21*, 3841-3851.

Xiang, S., Short, S.A., Wolfenden, R., and Carter, C.W. (1997). The structure of the cytidine deaminase-product complex provides evidence for efficient proton transfer and ground-state destabilization. Biochemistry *36*, 4768-4774.

Xie, Y., Blanc, V., Kerr, T.A., Kennedy, S., Luo, J., Newberry, E.P., and Davidson, N.O. (2009). Decreased Expression of Cholesterol 7{alpha}-Hydroxylase and Altered Bile Acid Metabolism in Apobec-1-/- Mice Lead to Increased Gallstone Susceptibility. J Biol Chem *284*, 16860-16871.

Xie, Y., Nassir, F., Luo, J., Buhman, K., and Davidson, N.O. (2003). Intestinal lipoprotein assembly in apobec-1-/- mice reveals subtle alterations in triglyceride secretion coupled with a shift to larger lipoproteins. Am J Physiol Gastrointest Liver Physiol *285*, G735-746.

Xu, M., Wells, K.S., and Emeson, R.B. (2006). Substrate-dependent contribution of double-stranded RNA-binding motifs to ADAR2 function. Mol Biol Cell *17*, 3211-3220.

XuFeng, R., Boyer, M.J., Shen, H., Li, Y., Yu, H., Gao, Y., Yang, Q., Wang, Q., and Cheng, T. (2009). ADAR1 is required for hematopoietic progenitor cell survival via RNA editing. Proc Natl Acad Sci USA *106*, 17763-17768.

Yamanaka, S., Poksay, K.S., Arnold, K.S., and Innerarity, T.L. (1997). A novel translational repressor mRNA is edited extensively in livers containing tumors caused by the transgene expression of the apoB mRNA-editing enzyme. Genes Dev *11*, 321-333.

Yamanaka, S., Poksay, K.S., Driscoll, D.M., and Innerarity, T.L. (1996). Hyperediting of multiple cytidines of apolipoprotein B mRNA by APOBEC-1 requires auxiliary protein(s) but not a mooring sequence motif. J Biol Chem *271*, 11506-11510.

Yang, and Smith, H.C. (1997). Multiple protein domains determine the cell type-specific nuclear distribution of the catalytic subunit required for apolipoprotein B mRNA editing. Proc Natl Acad Sci USA *94*, 13075-13080.

Yang, S.Y., Fugmann, S.D., and Schatz, D.G. (2006a). Control of gene conversion and somatic hypermutation by immunoglobulin promoter and enhancer sequences. J Exp Med *203*, 2919-2928.

Yang, S.Y., and Schatz, D.G. (2007). Targeting of AID-mediated sequence diversification by cis-acting determinants. Adv Immunol *94*, 109-125.

Yang, W., Chendrimada, T.P., Wang, Q., Higuchi, M., Seeburg, P.H., Shiekhattar, R., and Nishikura, K. (2006b). Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. Nat Struct Mol Biol *13*, 13-21.

Yu, X., Yu, Y., Liu, B., Luo, K., Kong, W., Mao, P., and Yu, X.-F. (2003). Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. Science *302*, 1056-1060.

Zahn, R.C., Schelp, I., Utermöhlen, O., and von Laer, D. (2007). A-to-G hypermutation in the genome of lymphocytic choriomeningitis virus. J Virol *81*, 457-464.

Zennou, V., and Bieniasz, P.D. (2006). Comparative analysis of the antiretroviral activity of APOBEC3G and APOBEC3F from primates. Virology *349*, 31-40.

Zhang, H., Yang, B., Pomerantz, R.J., Zhang, C., Arunachalam, S.C., and Gao, L. (2003). The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. Nature *424*, 94-98.

Zhang, J., and Webb, D.M. (2004). Rapid evolution of primate antiviral enzyme APOBEC3G. Hum Mol Genet *13*, 1785-1791.

Zhang, Z., and Carmichael, G.G. (2001). The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. Cell *106*, 465-475.